MACHINE LEARNING BASED CHRONIC KIDNEY DISEASE PREDICTION MODEL

HANA ENGIDASEW TAYE

OFFICE OF GRADUATE STUDIES

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

JULY, 2021

ADAMA, ETHIOPIA

Machine Learning Based Chronic Kidney Disease Prediction Model

Hana Engidasew Taye

A Thesis Submitted to

The Department of **Computer Science and Engineering**

School of Electrical Engineering and Computing.

Presented in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

July, 2021

Adama, Ethiopia

Machine Learning Based Chronic Kidney Disease Prediction Model

Hana Engidasew Taye

Advisor: Teklu Urgessa (PhD.)

A Thesis Submitted to

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing.

Presented in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

July, 2021

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled “Machine Learning Based Chronic Kidney Disease Prediction Model” is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of material that are used for this thesis have been duly acknowledged through citation

_____	_____	_____
Name of Student	Signature	Date

RECOMMENDATION

I certify that I have read and revised the dissertation entitled “Machine Learning Based Chronic Kidney Disease Prediction Model” prepared under our guidance by **Hana Engidasew** submitted in partial fulfillment of the requirements for the degree of Masters of Science. Therefore, I recommended the submission of revised version of that thesis to the department following the applicable procedures.

_____	_____	_____
Advisor	Signature	Date

APPROVAL SHEET

I, the advisor of the thesis entitled "Machine Learning Based Chronic Kidney Disease Prediction Model" by **Hana Engidasew,** hereby certified that the recommendation and suggestion made by the board of examiners are appropriately incorporated into the final version of the thesis.

_____	_____	_____
Advisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by **Hana Engidasew,** have read and evaluated the thesis entitled "Machine Learning Based Chronic Kidney Disease Prediction Model" and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree Master of Science in Computer Science and Engineering.

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC)

_____	_____	_____
Department Head	Signature	Date
_____	_____	_____
School Dean	Signature	Date
_____	_____	_____
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENTS

First and foremost, I would like to thank God for all his blessing from the beginning to the end of this research. I am extremely grateful to my advisor Dr. Teklu Urgessa for his friendly approach, invaluable advice, his patient, critical comments, and commitment to guide and continuous support that greatly improved this thesis work.

I am extremely grateful to Ato Endris Mohammed (PhD. Candidate at ASTU) for his friendly support, advice, appreciation and motivate to accomplish my thesis throughout my study. My brother Maireg Engidasew and my cousin Shemeles taddese for supporting in data collection from Thikur Anbesa specialized hospital by connecting different medical Doctors. I would like to thank all my friends and working staffs for their cherished time spent together, ideas, and support until the end of my thesis.

I would like tank Dr. Ayele Belachew HR and Finance Head and W/Ro Desta health information system Department Head at Tikure Anbesa specialized hospital for helping, allowing, and providing me valuable information from health system. My appreciation also goes out to my family, I am grateful to my mother Asnakech Assefa for supporting me and my husband shamble Nigatu, and friends for their encouragement and support all through my study.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LISTOFEQUATION

LIST OF ABBREVIATIONS

ACR	Albumin-to-creatinine ratio
AI	Artificial Intelligence
ANN	Artificial Neural Network
BMI	Body Mass Index
CAD	Coronary Artery Disease
CFS	Correlation based Feature Selection
CKD	Chronic Kidney Disease
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
COVID_19	Corona Virus
CPU	Central Preprocessing Unit
CRF	Chronic Renal Failure
CSV	Comma Separated Value
CV	Cross Validation
CVD	Cardiovascular Disease
DM	Diabetic Mellitus
DT	Decision Tree
ESRD	End Stage Renal Disease
FCBFS	Fast Correlation-Based Feature Selection
FN	False Negative
FP	False Positive
FS	Feature Selection
GFR	Glomerular Filtration Rate
HTN	Hypertension
IG	Information Gain
KDIGO	Kidney Disease Improving Global Outcome
KNN	K-Nearest Neighbors
LG	Logistic Regression
MDRD	Modification Of Diet In Renal Disease
ML	Machine Learning

MLP	Multilayer Perceptron
NB	Naïve Bayes
NICE	National Institute for Health and Care Excellence
NCD	Non-Communicable Diseases
NKF-KDOQI	National Kidney Foundation's Kidney Disease Quality Outcome Initiative
PNN	Probabilistic Neural Network
RBC	Red Blood Cell
RBF	Radial Basis Function
RF	Random Forest
RRT	Renal Replacement Therapy
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UCI	University of California, Irvine
WBC	White Blood Cell
WHO	World Health Organization

ABSTRACT

Chronic kidney disease was the most and primary cause of death internationally. If people build up CKD then their kidney become damage and over time may not clean the blood like as healthy kidney. According to WHO, most of developing countries like Ethiopia CKD is a growing problem and majority of patients with kidney disease die because lack of proper handling or give treatments. ML plays a huge role in healthcare system and it can effectively help and assist on decision support on medical centers. The main purposes of this study develop and recommend a machine learning algorithm for prediction of CKD. The CKD dataset collected from TASH which has 2135 with huge number of missing value and then after effectively preprocessing the dataset we use four types of machine learning algorithms, these are support vector machine, Decision tree, Random forest and multi- layer perceptron Artificial neural network. From those machine learning algorithms, MLP ANN the best performance with 99% accuracy.

Key Words: Chronic kidney disease, Tikure Anbesa Specialized Hospital,World Health Organization, Artificial Neural network, Multi-Layer Perceptron.

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Now days we are witnessing that, kidney disease are most critical healthcare problem for different age groups of the population[1]. This disease may go through different stages to reach an acute kidney or chronic level. On the other hand, kidney problems can also develop suddenly (acute) or over the long term (chronic) [2]. Both AKI and CKD enforce major burden on healthcare system but receive less attention than other non-communicable disease on the global policy of agenda[3]. The chronic kidney disease define with the abnormalities of kidney function that identified by following the patient for greater than three months and including with other implications of health complications[3].

In the process of detecting and diagnosing nephrology is used the term chronic kidney disease or chronic kidney failure, interchangeably while describing the gradual loss of kidney function[4]. Different researches states that, chronic kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population[5]. CKD is a significant public health problem worldwide, especially for low and medium-income countries[5]. The prevalence of CKD in the general population is 13.4% globally, while pooled prevalence of CKD is 10.1%in the general population, 24.7% in hypertensive, and 16.6% among diabetes mellitus patients in Africa [6].

Among the cause that leads to CKD, Hypertension, diabetes, and aging are considered a major leading factors. Therefore, features of those diseases that most determine for CKD have to be investigated while making decision[6]. The sub-factors such as high blood pressure, coronary artery disease, and anemia, bacteria and albumin in the urine, complications of some drugs, sodium and potassium deficiency in the blood and a family history[2][7]. Patients with CKD suffer from various side effects such as complications include damage to the nervous and immune systems that disrupt daily activities[8]. In the early stages of chronic kidney disease, it may have few signs and symptoms. CKD does not show obvious symptoms in its early stages and the disease may not be detected until the kidney loses about 25% of its function [5][9][10].

A well-accepted CKD screening test is the glomerular filtration rate (GFR), measuring the kidney to estimate its function status regarding the glomerular[11]. CKD is classified into Stages I–V according to the estimated glomerular filtration rate (GFR)[12]. GFR is estimated having mathematical equations using serum creatinine, age, sex, body size, ethnic origin [9][11][13][14][15].

In this regard, the increasing alarm of CKD shows more work have to be made in diagnosis and detecting process to enhance the decision accuracy. Different research states that Kidney disease is very dangerous if not immediately treated on time, and may be fatal. If the doctors have a good tool that can identify patients who are likely to have kidney disease in advance, they can heal the patients in time [7]. Moreover, the diagnosis and detection process of CKD mainly relied on nephrologist's knowledge and tools; which is error prone[16]. Currently, AI and machine learning algorithms used in different areas to enhance the decision making process, among those areas healthcare service is taking the advantages of it[17]. The wide application of ML in the medical field helps to promote medical innovation, reduce medical costs, and improve medical quality. Some representative works in this area such as cancer prognosis and prediction[18], 1-year mortality prediction model after discharge in clinical patients with acute coronary syndrome[19], heart disease prediction[20] Some representative works in this areas are machine learning methodology for diagnosing CKD[5], Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms [2], The existence of chronic kidney disease by imposing various classification algorithms on the patient medical record[12]. With machine learning techniques, it is possible to analyze lab records and other information off patients for the early detection of CKD[7].

This paper aims in builds a model for prediction in CKD using our county health care data. The data was collected from Tikure Anbesa specialized Hospital that found in Addis Abeba Ethiopia. Tikure Anbesa hospital Have Health care information system that can manage patient data, renal clinic is one of those Tikure Anbesa specialized hospital clinics. Based on the data identify the attributes to define different stages of CKD and classify the patient in which class (CKD/ Not CKD) of kidney disease

1.2 Motivation of the study

Kidney disease is a major public health challenge worldwide [21]. Especially in our country life is so expensive and monthly or daily incomes of people are low so that patient was not going to hospitals and clinics; they can simply take painkiller as well as most hospitals and clinics are not ready to handle Kidney problems early. According to WHO, "If risk Factors are identified early, acute kidney injury and chronic kidney disease can be prevented and, if kidney disease is diagnosed early, worsening of kidney function can be slowed or prevent by inexpensive intervention"[21]. Now a day's kidney problem news is published in social media for the purpose of collecting money to individuals for dialysis and kidney transplant due to this I motivated for this study is thinking of using Machine learning prediction analysis to develop a prediction model which to identify early stage and progression of chronic kidney disease which helps supporting in decision making of health professional to prevent the CKD problem.

1.3 Statement of the Problem

Non communicable diseases are the most common cause of premature death and from those non-communicable diseases; CKD is take the first place [22]. Chronic Kidney Disease is a main challenge for health care system around worldwide[23].People with CKD, there must be focus on applying proven, cost effective treatments for many people as much as possible, taking into account limited needs, human and economical recourses.

Currently chronic kidney disease (CKD) makes a huge damage on societies with increasing in alarming rate[24]. Different efforts have been made to advance early treatment to prevent the growth of its stage before reaches to chronic disease[3]. Recent studies suggest that some of the bad outcome can be prevented using early detection and treatment. In this regard, various AI and Machine-Learning-Based diagnosing techniques and approaches such as among them KNN, RF, LR, SVM and ANN have been used for CKD detection. Most studies use online dataset which is collected from university of California Irvine (UCI). As gap they put on their study relatively use small dataset which contain total of 400 which have two classifications (CKD and NOT CKD) with more missing value and suggest that real time dataset was collected with more patient is more preferable for achieving highest accuracy. So that this study will collect and prepare a chronic kidney dataset from local hospitals to predict chronic kidney disease using machine

learning algorithm with high volume of dataset record. In addition to that, the limited data sets used in the previous studies have a negative impact on accuracy and performance of the diagnosing system. Therefore, this study could help physicians, health professionals to detect the disease and give fast treatment for the patient in a short time.

1.4 Research questions

Based on the above problem, this study will have answered the following questions

- What is the state of CKD problem and Solution Using machine learning?
- Which machine learning Model is better to predict chronic kidney disease?
- How to measure the performance of the Model?

1.5 Objective of the study

1.5.1 General Objective

The general objective of this study is to investigate, design and develop Machine learning based prediction model for CKD

1.5.2 Specific objectives

To achieve the general objective of this study the sub objectives are listed as follows

- To review literature on the concepts of Machine learning and CKD prediction Model
- To collect necessary data related to problem and extract features
- To design model and represent the knowledge of the area
- To develop prediction model for chronic kidney disease
- To evaluate and validate the performance of the Model

1.6 Scope and Limitation of the study

1.6.1 Scope of the study

There is different ways of developing the disease prediction models. This research focuses on using Machine learning classifications algorithms to identify the patient having CKD or not CKD. This helps the healthcare specialist to give appropriate decisions at the time of patient diagnosis. The dataset was collected from local hospital that was Tikure Anbesa Specialized Hospital and the dataset helps to measure the nonexistence or existence of the CKD. The dataset has 21 features and used SVM, DT, RF and ANN machine learning classification algorism are used to develop a model.

1.6.2 Limitation of the study

This research passes several limitations on different stages of the research development process. In Ethiopia, within the medical system it was difficult to collect the dataset with several reasons. One of the reasons was patient history confidentiality, they cannot easily allow the patient history to the external user and they need Ethical clearance to assure the patient privacy. In Tikure Anbesa Specialized Hospital Data maintains using Health care electronics data system but they have not organized in the way this study needed the dataset and it is time-consuming to convert their data in to appropriate dataset format. Data collection process almost takes one year due to COVID-19, Ethical Clarence and other several reasons. Additionally, this study performs arrange of the dataset in the form of online Machine Learning Chronic Kidney Dataset. This process consumes a lot of time and this reduces the working time of the study.

1.7 Significant of the study

CKD is not only an individual problem but also family and Environmental problem. Therefore, to identify CKD at early stage and its Risk Factors help to undertake the problem. This disease especially if the patient reaches end stage of renal diseases (ESRD), this challenge leads to family to face more problems due to dialysis fee or kidney transplant. The result of this study will provide several importance's for different groups. The developed model will assist the physicians by providing necessary facts and they can get accurate prediction of CKD to prevents

the problem easily. And also patient beneficiary because the physician effectively uses the system therefore Mortality rate of patient is reduced, easily prevent the problem and improve patient outcomes and reduce patient cost for dialysis. In Health centers and Hospitals side, the result of this study will reduce the cost of health resources and budget by supporting accurate decision making for disease and reduce the total budget for health centers. This model helps the medical institution can improve their clinical settings; also early prediction strategy can prevent the development of CKD and minimize its clinical impact in our country health centers. In system developer side, the result of this study will be used as knowledge gaining techniques in the problem area and will help as a corner stone for future research on this area.

1.8 Organization of the Thesis

This Paper organized into six chapters.

Chapter 1 Introduction: describes the background of the study and motivation, the statement of problems and research Questions, the objective of the study, the scope and limitation of the study, and significance of the study, and the organization of the study.

Chapter 2 Literature Review and Related Works: provide relevant information about CKD and its current status, the algorithms used to predict CKD problems, machine learning algorithms for classification of CKD and related works.

Chapter 3 Methodology: This chapter discussed the research methodology such as data collection and data preprocessing techniques, feature selection and machine learning algorithms with prediction model evaluation techniques.

Chapter 4 System Modeling and Experimental setup: this chapter discusses the experimental setup of chronic kidney disease prediction model description.

Chapter 5 Implementation and Discussion: this chapter discusses the implementation of the prediction model and Discussion of the result of this study.

Chapter 6 Conclusion and future work: concludes the study and recommendation of feature research.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

On this chapter, the study will review necessary concepts about chronic kidney disease identification mechanisms and machine learning technology that help the health care system in early prediction of CKD. And review several related papers to help this study further explore the research problem and how the Machine learning detect the CKD in early stage.

2.1 Review concepts

2.1.1 Chronic Kidney Disease

The two kidneys are located to the rear of the abdominal cavity on either side of the spine. They normally weigh about 5 once each, but receive about 20% of the blood flow coming from the heart. The urine produced by each kidney drains through a separate urethra into the urinary bladder, located in the pelvic region. In the human body kidney is main and important organ which used to manage fluid levels, electrolyte balance, and other factors that keep the internal environment of the body consistent and comfortable. Kidney diseases are disorders that affect the functions of the kidney. During the late stages, kidney diseases can cause kidney failure. Kidney diseases are disorders that affect the functions of the kidney. Kidneys can get damaged that means they cannot do all the things they should. This is called chronic kidney disease or CKD. Chronic kidney disease can affect anyone.

In medical science nephrologists uses mostly two major tests are used for detecting CKD. A blood test that checks the glomerular filtration Rate (GFR) and Urine test to check for albumin [9]. Factors that can affect CKD are genetics, hypertension, diabetes, obesity, aging and others. The international kidney disease development guidelines and standard foundations such as US National Kidney Foundation Kidney Disease Outcomes Quality Initiative (KDOQI) and KDIGO (Kidney Disease Improving Global Outcome), describe important information and developments about CKD. According to the KDIGO CKD and English National Institute for Health and Care Excellence (NICE) CKD guidelines guiding principle the kidney patient recognized by two tests, these are the blood test tests to check, how well the kidneys filter the blood to eliminate

creatinine, a normal muscle breakdown waste. In comparison, the urine test will indicate that protein remains in the urine and in particular, protein (albumin) is a blood component that is usually not transmitted into the urine by the kidney filter. When the urine test indicates that albumin occurs, it means that the kidney filters are damaged and can represent chronic kidney disease. Chronic Kidney Disease (CKD) is defined as kidney damage or glomerular filtration rate (GFR) <60 ml/min/1.73 m2 for more than 3 months with implications for health.

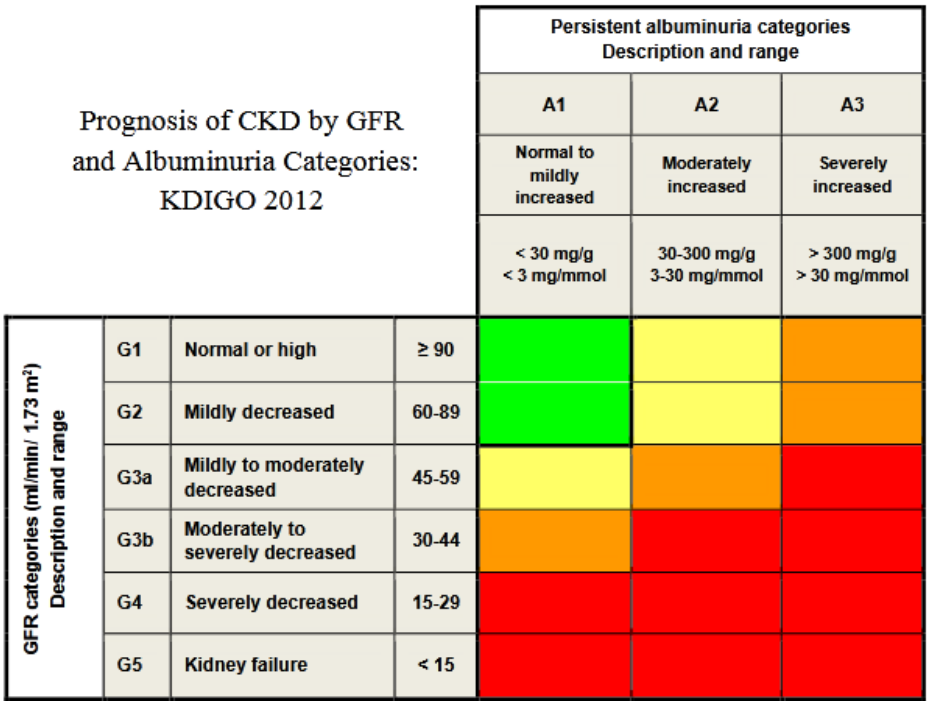


Figure 2.1GFR and Albuminurea categories KDIGO 2012[25]

Fig. 2 shows the Prediction of CKD by GFR and albuminuria categories. Green, low risk of disease progression; yellow, moderately increased risk of disease progression; orange, high risk of disease progression; red, very high risk of disease progression. CKD chronic kidney disease, GFR (glomerular filtration rate) and ACR (albumin-to-creatinine ratio)

2.1.2 Machine learning

Machine learning (ML) is a kind of artificial intelligence (AI). The core of it is algorithmic methods, which enable the machine to solve problems without specific computer programming. The wide application of ML in the medical field helps to promote medical innovation, reduce

medical costs, and improve medical quality. However, related research on solving clinical problems through ML in nephrology still needs to increase. Understanding the purpose and method of ML application and the current situation of its application in nephrology is a prerequisite for correctly addressing and overcoming these challenges. Machine learning has already been used to detect human body status, analyze the relevant factors of the disease, and diagnose various diseases. For example, the models built by machine learning algorithms were used to diagnose heart disease, diabetes and retinopathy, acute kidney injury, cancer and other diseases.

The term machine learning (ML) is extremely hot nowadays, and a number of clinical prediction model studies used this sort of technologies [26]. While the healthcare sector is being transformed by the ability to record massive amounts of information about individual patients, the enormous volume of data being collected is impossible for human beings to analyze[26]. Machine learning provides a way to automatically find patterns and reason about data, which enables healthcare professionals to move to personalized care known as precision medicine. [1]. The combination of machine learning, health informatics and predictive analytics offers opportunities to improve healthcare processes transform clinical decision support tools and help improve patient outcomes[27]. The promise of machine learning's changing healthcare lies in its ability to leverage health informatics to predict health outcomes through predictive analytics, leading to more accurate diagnosis and treatment and improving physician insights for personalized and cohort treatments.

ML technology can be broadly divided into supervised learning, unsupervised learning, semi-supervised and reinforcement learning according to different modeling needs.

2.1.2.1 Supervised Learning

Supervised Learning algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables) and With these set of variables, we generate a function that map inputs to desired outputs[28][29]. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc[28][29].

2.1.2.2 Unsupervised Learning

In unsupervised learning algorithm, we do not have any target or outcome variable to forecast or estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific interference and other areas including probability density estimation, finding association among features, and dimensionality reduction[28][29][9]. Examples of Unsupervised Learning: Apriori algorithm, K-means

2.1.2.3 Semi-supervised Learning

Semi-supervised learning algorithm is defined by both supervised and unsupervised learning having both labeled and unlabeled data, and jointly learns knowledge from them. Labeled data is basic information that has meaningful marker so that the algorithm can understand the data, at the same time as unlabeled data lacks that information. By using this combination, machine learning algorithms can learn the unlabeled data[9]

2.1.2.4 Reinforcement Learning

In Reinforcement learning algorithm, the machine is trained to make specific decisions. By making the machine to exposed an environment where it trains itself continually using trial and error, this machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions[28][29][9]. Reinforcement learning is used to resolve problems of decision making (usually a sequence of decisions), such as robot perception and movement, automatic chess player, and automatic vehicle driving [28]. Example of Reinforcement Learning: Markov Decision Process

Most researchers use classification techniques in disease prediction and diagnosis. Most of the prior studies that we reviewed for this research was Decision Tree (DT), Random Forest (RF),Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Logistic Regression (LR), Artificial-Neural Network (ANN) and the majority uses more than one classification algorithm to develop CKD prediction model[2][9][8][11][12][15][30][31][32]

The following table will show frequently used classifier algorithms for chronic kidney disease prediction

Table 2.1Machine learning classification algorithms and their pros and cons

Classification Algorithm	advantage	Disadvantage
Naïve Bayes Classifier algorithm	Simple, Easy and fast Not Sensitive to irrelevant features Work grate in practice Need Less Training Data For both multi-class and binary classification Work with continuous and discrete data	Accepts every feature is independent. This isn't always the truth
Decision Tree	Easy to understand Easy to generate rules There are almost null hyper-parameters to be tuned Complex decisions tree models can be significantly simplified by its visualization	Might be suffer from overfitting Does not easily work with non-numerical data Low prediction accuracy for dataset in comparison with other algorithms When there are many class labels calculation can be complex
Support Vector Machine	Fast algorithms Effective in high dimensional space Great accuracy Power and flexibility from Kernels Works very well with clear margin of separation Many application	Doesn't perform well with large data set Not so simple to program Doesn't perform well when the data comes with more noise which means target classes are overlapping

Random Forest	The over fitting problem does not exist Can be used for feature engineering i.e for identifying the most important features among all available Run very well on Large Dataset Extremely flexible and have very high accuracy No need for preparation of the input data	Complexity Requires a lot of computational resources Time consuming Needs to choose the number of trees
KNN algorism	Simple to understand and easily to implement Zero to Little training time Works easily with multicast dataset Has good predictive power Do well in practice	Computationally expensive testing phase Can have skewed class distribution The accuracy can be decrease when it comes to high dimension data Need to define a value for parameter k
Neural Network	for both regression and classification problems Used ranging from voice recognition up to cancer prediction good to model with nonlinear data with large number of inputs It works by splitting the problem of classification into a layered network of simpler	Neural networks are black boxes, meaning we cannot know how much each independent variable is influencing the dependent variables. It is computationally very expensive and time consuming to train with traditional CPUs Neural networks depend a lot on training data

	elements Once trained, the predictions are pretty fast can be trained with any number of inputs and layers work best with more data points	

2.1.2.5 Neural Network in CKD prediction Model

A neural network is defined as a parallel-distributed processor with high-computation power made of various processing units, called neurons[33]. NN is a novel process of programming computers. It is represented of the human brain's information processing mechanism. NN is applied on more applications such as pattern recognition, diagnosis of diseases and data classification, through a learning process [27]. NN has many types of networks such as feed-forward network and back-propagation network. NN is composed of Input Layer which is the input units represent the raw data that is fed into the network, Hidden Layer that is the hidden unit is specified by the input units and the weights on the connections between the input and the hidden units of the Output Layer - The attitude of the output units depends on the activity of the hidden units and the weights between the hidden and output units [27].

This study aim to use several supervised machine learning algorithms to develop a prediction model, the algorithms include Decision tree, Random forest, Support vector Machine Artificial Neural network to predict chronic kidney disease.

2.2 Related Works

There are many researchers who work on prediction model for CKD using different machine learning algorithms.

JIONGMING QIN1, et al. [5].proposed machine learning methodology for diagnosing CKD. Dataset obtained from university of California Irvine (UCI) machine learning repository which has large number of missing values, KNN imputation was used to fill in the missing values, which select several complete samples with the most similar measurement to process the missing data for each incomplete dataset, they use six machine learning algorithms such as logistic

regression, random forest ,support vectors machine, K- nearest neighbor, naive bays classification and feed forward neural network were used to establish model. Among these machine learning models random forest achieve the best performance with 99.75% diagnostic accuracy and they analyze the misjudgment generated by the establish model to propose an integrated model that combines logistic regression with random forest by using perception and achieve average accuracy of 99.83%

N.A. Almansour, et al.[8]aim to assist in the prevention of CKD by utilizing machine learning techniques to diagnose CKD at an early stage, they focus on applying different machine learning classification algorithms to a dataset of 400 patients and 24 attributes related to diagnose of CKD. They use Artificial Neural Network and support vector machine as classification techniques and all missing value in the dataset were replaced by the means of the corresponding attributes. The final model of the two proposed techniques were developed using the best obtained parameters and features. The empirical result from experiment indicated that ANN performed better than SVM with accuracy of 99.75% and 97.75% respectively

Zixian Wang, et al[7] this study analyzes CKD using machine learning technique based on dataset from UCI machine learning dataset warehouse. CKD is detected using the apriori association techniques for 400 instance of chronic kidney patients with 10-fold-cross-validation test and results are compared across a number of classification algorithms including ZeroR, OneR, naïve Bayes, J48 and IBK (K-nearest Neighbor). The dataset preprocessed by completing and normalizing missing data. They select the most relevant feature from the dataset for improved accuracy and reduce training time. The result is 99% detection accuracy for CKD based on Apriori

ShanilaYunusYashfi, et al.[2]they analyze data of CKD patient and proposed a system from which it will be possible to predict the risk of CKD. They should have used 456 patient's data which is collected from online UCI machine learning repository and real time dataset from Khulna city medical college. Python high level interpreted programming language used to develop the system. They trained the data using 10-fold CV and applied random forest and ANN. The accuracy achieved by random forest algorithm is 97.12% and ANN is 94.5%.

H. Kriplani et al. [1]. They propose deep neural network which predict the presence or absence of CKD with an accuracy of 97%. They use UCI Dataset form ML Repository. This dataset has been collected from the Apollo hospital (Tamilnadu) nearly 2 months of period and has 25 attributes, 11 numeric and 14 nominal. Total 400 instances of the dataset in which 224 of them are used for the training to prediction algorithms

N V GanapathiRaju et al [12]. This research work primarily concentrated on finding the best suitable classification algorithm which can be used for diagnosis of CKD based on the classification report and performance. They have compared the performance of five classifiers in the prognosis of CKD. The experimental results of our proposed method have demonstrated that RF and XGB have produced superior prediction performance in terms of classification accuracy for our considered dataset.

SiddheshwarTekale et al [15]. They have studied different machine learning algorithms. By using data of CKD patient with 14 attributes and 400 records and apply ML algorithms like decision tree and support vector machine. Therefore, from their result analysis Decision tree algorithm gives 91.75% and Support Vector Machine gives accuracy of 96.7.

Table 2.2. Summary of literature Review

SN	Author	Main Idea	Method and Material (Dataset)	Feature	Tools and Technology	Remark
1	JIONGMING QIN1, et al. [5]	A ML Methodology for diagnosing CKD	• UCI(400) samples(2015) • 250 samples belongs to CKD and 150 samples belongs to NOT CKD	• 24 predictive variables or feature (11 numerical variables, 13 nominal Variables and 1 class variables	ML Algorithms such as KNN, Logistic Regression, Random Forest, SVM, Naïve Bays, Feed Forward Neural Network	•The available data sample is relatively small •There are only two categories CKD/NOT CKD and cannot diagnoses the severity of CKD •In the future large no of more complex data will be collected to train the model and to detect the severity CKD
2	N.A. Almansour, et al.[8]	To assist in the prevention of CKD by utilizing ML techniques to diagnosis CKD at early stages	UCI(400 patient) samples(2015)	24 predictive variables or feature (11 numerical variables, 13 nominal Variables and 1 class variables	ML Algorithms (ANN and SVM)	the need to use deep learning as the starting point a slightly less necessary at the beginning of this study because one of the principles of machine learning is to start with simple classifiers and then gradually proceed with complex classifiers depending on the need
3	Zixian Wang, et al [7]	ML based Prediction system for CKD using Associative Classification Techniques	• UCI (400) samples (2015) 250 samples belongs to CKD and 150 samples belongs to NOT CKD	• 16 selected attributes is used	• WEKA tool for selection features • ZeroR, OneR, naive Bayes, J48,andIBk (k-nearest-neighbor).	The future research should analyze different supervised and unsupervised ML techniques with additional performance metrics for the better CKD prediction
4	ShanilaYunusYashfi, et al. [2]	Risk Prediction Of Chronic Kidney Disease Using Machine Learning	455 patients' data. Online data set which is collected from UCI Machine Learning Repository and real	• 25 features	• Python as a high-level interpreted programming language for developing our	use wrapper method to extract significant features more accurately. Most of our data are from online. In future we intend to work with real time

		Algorithms	time dataset which is collected from Khulna City Medical College are used.		system. • trained the data using 10-fold CV and applied Random forest and ANN	dataset and achieve higher accuracy
5	H. Kriplani et al. [1]	Prediction of Chronic Kidney Diseases Using Deep Artificial Neural Network Technique	• UCI (400) samples (2015) uses only 224 patient dataset 105 samples belongs to CKD and 119 samples belongs to NOT CKD	• 25 attributes	• DNN, Naïve Bayes, Logistic Regression, Random forest, Adaboost and SVM	The resultsWould be more promising with increase in dataset. Thus it has outperformed other classifiers and is able to detect the chronic kidney disease more efficiently
6	N V Ganapathi Raju et al [12]	ascertain the existence of chronic kidney disease by imposing various classification algorithms on the patient medical record	UCI(400 patient) samples(2015)	• 25 attributes	Support Vector Machine, Random Forest, XGBoost, Logistic Regression, Neural networks, Naive Bayes Classifier	For the future, we are working enhancing the performance of prediction system accuracy by ensemble different classifier algorithms
7	Siddhesh warTekale et al [15]	Prediction of Chronic Kidney Disease Using Machine Learning Algorithm	UCI(400 patient) samples(2015)	14 selected attributes is used	Decision Tree, SVM	The strength of the data is not higher because of the size of the data set and the missing attribute values

CHAPTER THREE

RESEARCH METHODOLOGY

At this part of our study we discussed the methodology of the research and every process that include in chronic kidney disease prediction model development, these activities performed using machine learning algorithms.

In healthcare system best diagnosis tool is the brain of doctor. But if lab and diagnosis records analyze using electrical information system then the doctors make better decision about patient diagnosis and treatment options. The value of machine learning in health care system, it can analyze and process huge dataset that is beyond the scope of human mind capability and then reliable convert analysis of that huge data into clinical insights, then it assist medical doctors in planning and providing care, altimetry leading to improved outcomes, lower cost of health care's and increase patient satisfaction.

In this study several methods, procedures and techniques have to be used to build a machine learning models. At this chapter, first we discussed about the data collection are explained, second we make clear the data preprocessing methods then finally we give details the machine learning algorithms for developing a CKD prediction model and explain the performance evaluation methods are discussed.

The research methodology shows in the following Diagram

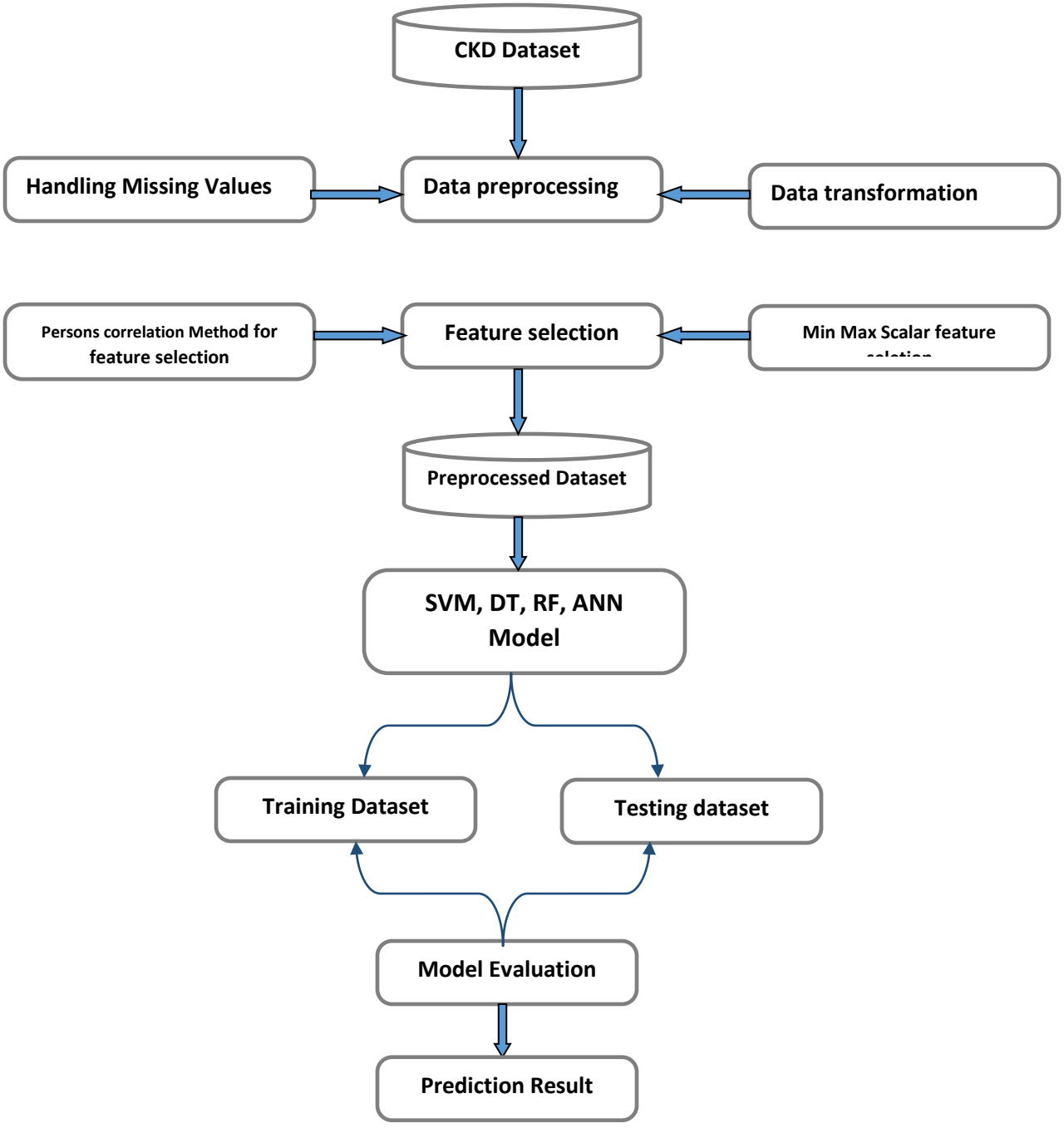


Figure 3.1Methodology for proposed system

3.1 Data Collection

The data to conduct this research was collected from Tikure Anbesa specialized hospital (TASH). It is a public hospital found in Addis Ababa, the capital of Ethiopia. TASH was established in 1972 and is affiliated with Addis Ababa University. It is the largest referral hospital in the country, after 1998 it was transferred to the University teaching hospital by the federal ministry of health. Currently TASH is the main teaching hospital and it has 200 doctors for several clinics. Recently TASH establish the Electrical Health information system(EHIS) to hold and maintain the patient history data. This Electrical Health information system contains several patient data and they give us 2 Years Chronic Kidney disease patent records. The chronic kidney disease dataset contains 2135 records and 20 features and one class.

3.2 Chronic Kidney Disease Prediction Modeling

3.2.1 Data Preprocessing

Data is the most important part of all data analytics, machine learning, and artificial intelligence. Data in the real world is "dirty" which means data is incomplete, noisy and inconsistent, inaccurate (containing errors and outliers) no quality data which lead our model to no quality mining result. Data preparation can make or break the prediction ability of our model. Preprocessing of data can lead to unexpected improvement in model accuracy.

In our chronic kidney dataset there are missing values and outliers so we should perform data cleaning, handling missing values for categorical and numerical values, converting nominal values to number values, feature selection and other preprocessing tasks. Our study follows the following data preprocessing Architecture to clean our dataset.

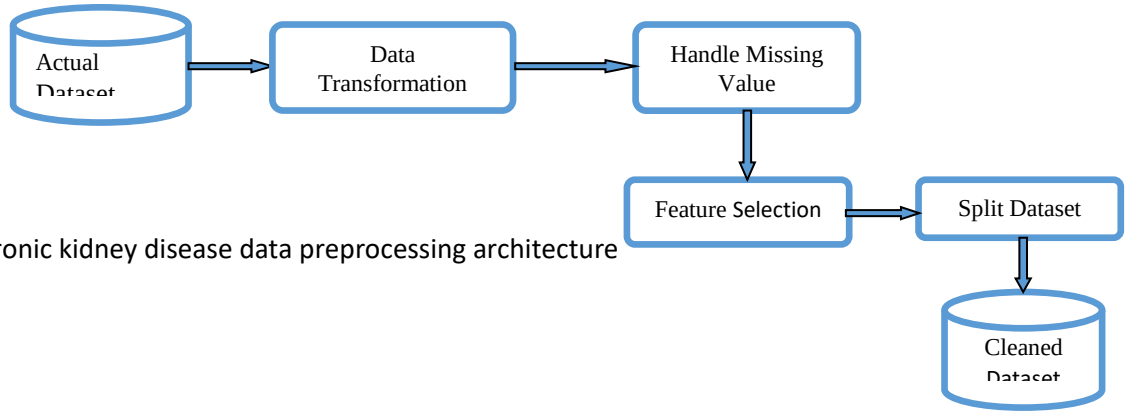


Figure 3.2 Chronic kidney disease data preprocessing architecture

3.2.1.1 Data Description

The dataset used for this study contains 2135 records and each record have 20 features and one categorical features (class), 15 are numerical values, 5 have nominal values and 1 categorical values. The dataset describes in the following tables.

Table 3.1. Dataset Features Description

Symbols	Features full name	class	Value in range
Age	Age	Numeric	Age (number)
Gender	Gender	Nominal	M and F
Bp	Blood pressure	Numeric	In mm/Hg
Sg	Specific gravity	Numeric	(1.005,1.010,1.015,1.020,1.025)
Al	Albumin	Numerical	(0,1,2,3,4,5)
Chl	Chloride	Numerical	in mEq/L
Sod	Sodium	Numeric	in mEq/L
pcv	Packed cell volume	Numerical	in cells/cumm
Pot	Potassium	Numeric	in mEq/L
Bun	Blood Urea Nitrogen	Numeric	in mgs/dl
Scr	Serum Creatinine	Numeric	in mgs/dl
Hgb	Hemoglobin	Numeric	in gms
Rbcc	Red blood cell count	Numeric	in millions/cmm
Wbcc	White blood cell count	Numeric	in cells/cumm
Mcv	Mean cell volume	Numeric	in cells/cumm
Pltc	Platelet count	Numeric	in mEq/L

Htn	Hypertension	Nominal	yes,no
Dm	Diabetes Mellitus	Nominal	yes,no
Ane	Anemia	Nominal	yes,no
Hd	Heart disease	Nominal	yes,no
Classification	Class	Nominal	ckd,notckd

3.2.1.2 Data Cleaning

The data cleaning process is making the real world data correct and consistent. Machine learning ("ML") algorithms require a systematic data structure in which every data sample should follow the same structure. While the human brain would just ignore missing data during the learning process, algorithms cannot process data with even minor inconsistencies.

3.2.1.3 Data Transformation

Our dataset contains some attributes with Nominal data, some attributes with Numerical data machine learning models only manipulate with numbers. So we transform all of the nominal data are converted into numerical data. For example, Hypertension (htn) Cell values: Yes = 1; No = 0. Diabetics Mellitus (DM) cell Values: Yes = 1, No = 0. Gender value: M=1, F=0. Anemia (Ane): Yes = 1, No = 0. Heart disease (Hd) cell values: Yes = 1, No = 0. And categorical class feature also converted CKD=1 and notCKD=0.

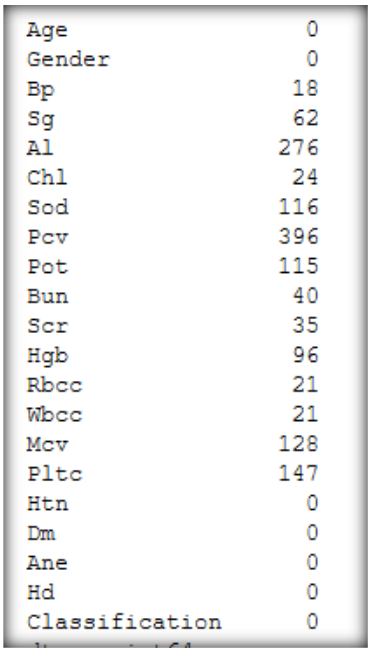
3.2.1.4 Handling Missing value

Missing and/or corrupted data often contaminates Medical data.[34] It occurs when there is no stored data value for the variable in an experiment or test. The absence of data is a common occurrence that can significantly affect the conclusions drawn from the data.

Medical data often contains many missing values, which makes analysis difficult for researchers who want to build a model using those data. Medical data have missing values for various reasons; for example, due to personal privacy, it is often difficult to collect complete data [35]. Although hospital systems are able to capture the overall data

measurement results, some patient data will remain missing from the database. Another reason for this is that some doctors usually write important diagnostic information in a easy text form, which is not easily converted into a machine-readable format [35]. These limitations make it difficult for algorithms to capture patterns in medical datasets.

Our dataset is created from the data gathered from patient history data, which contains missing values and the missing data can't be deleted because we miss necessary information that used in this study. And most features have missing values as shown Figure 3.3 because of this we cannot delete the features that have missing values.



Age	0
Gender	0
Bp	18
Sg	62
Al	276
Chl	24
Sod	116
Pcv	396
Pot	115
Bun	40
Scr	35
Hgb	96
Rbcc	21
Wbcc	21
Mcv	128
Pltc	147
Htn	0
Dm	0
Ane	0
Hd	0
Classification	0

Figure 3.3The number of missing value in each feature.

This study uses Imputation method which is the process of replacing the missing data with approximate values. Instead of deleting any columns or rows that has any missing value, this approach preserves all cases by replacing the missing data with the value estimated by other available information. From imputation method this study will use simple imputer method which is used to impute or replace both numerical and categorical features missing value. This method uses mean, median, most frequent and constant statistical methods for numerical missing value and for categorical features use a strategy such as most frequent and constant method is used.

3.1.1.1 Feature Scaling

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data orit makes features in the same standing. Feature scaling have a major impact on some machine learning algorithms, these include SVM, KNN and NN[10], then we need to scale features before using in any model development. In health care system the features of the dataset may vary in terms of magnitude, range and unit and then Most of ML algorithms work on the magnitude of the measurement not on the unit[36]. We have done feature scaling using min-max scalar algorithms. The min-max algorithm makes features value between [-1, 1] or [0, 1] in fixed range.

$$\mathbf{x'} = \frac{\mathbf{X - Xmin}}{\mathbf{Xmax - Xmin}} \qquad \text{Equation 3. 1}$$

Where:

x’ = new min-max value

X = old value

Xmin = minimum value

Xmax = maximum value

3.1.1.2 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model [35].At this stage the study will be identifying the related features from a set of data and removing the irrelevant or less important features with do not contribute much to our target variable in order to achieve better accuracy for our model. Several Researchers uses different types of feature selection method these include Filter method, Wrapper method and Embedded Method.

One of the measures used for feature selection is dependency measures. Many dependencies based methods have been proposed. The main measure is Correlation based method. Pearson’s Correlation method is used for finding the association between the

independent features and the dependent feature. This study uses correlation method for feature selection. Correlation Method evaluates all attributes related to the target class via Pearson's correlation method, which provides a ranking of the attributes from high to low and this reduces processing time and the data dimension[37].

Figure 3.3Architecture of Feature selection

3.1.2 Machine Learning Models

This study uses Decision Tree classifier, Random Forest Classifier, Support Vector Machine and Artificial Neural Network to predict chronic kidney disease. Among form these algorithms try to establish our prediction model and we select better performance by analyzing their accuracy.

3.1.2.1 Support Vector Machine

This algorithm is the highest well-known and important supervised machine-learning algorithm that is works on classification and regression problems however mostly used for classification problem. SVM used kernel function to separate labeled data. One of the advantages associated with using kernels in SVM is that SVM applies kernel dentitions to non-vector inputs and Kernels can be defined based on a combination of different data types [11]. In SVM algorithm data classify into two data points in which hyper-plane lies between two branch. SVM creates a discrete hyper plane in the signifier space of the training data and compounds are classified based on the side of the hyper plane located[38].

SVM has been utilized in several fields, including bioinformatics and handwritten recognition[8]. SVM is also used in other applications, including medical diagnosis,

weather prediction, stock market analysis, and image processing. Similar to all other machine learning techniques, SVM is a computational algorithm that learns from experience and examples to allocate labels to objects[8].

3.1.2.2 Decision Tree

Decision tree algorithm belongs to supervised machine learning algorithm and used to solve both classification and regression problems, but mostly used for classification problems. Decision Tree algorithm resolve the classification problem by converting the dataset into a tree representation through sorting them by feature values. In decision tree every node indicates features in an instance to be classified and every leaf node indicates a class label the instance belongs to.

3.1.2.3 Random Forest

Random forest algorithm is an easy and flexible supervised machine learning algorithm that consists of various collections of decision trees and used for both classification and regression problems. Random forest works by building the forest which is an ensemble of decision tree, usually trained with the bagging method[39].. The bagging methods means a method of combination of learning model increase the overall results. This model consists of several decision trees and outputs the class target that is the highest selection results of the target output by each tree[40].

3.1.2.4 Artificial Neural Network

Artificial Neural Networks which is also known as ANN is basically a computational model[8]. These computing Models are a complex network of basic components or set of nodes called neurons[9]. The nodes are interconnected with each other. Each connection between the nodes has a specific weight associated with it. The basic structure of any neural network consists of three layers. The first layer is called the input layer and the middle layer called the hidden layer and the final layer called the output layer. Every neural network will definitely have all these three layers with one or multiple nodes. A simple and basic neural network will have one hidden layer. But a complex neural network can have several hidden layers[8]. The inputs to the network are fed through the initial input layers. The computation is done in the hidden layers and the respective

output is popped out through the output layers. They are many different kinds of neural networks available. Now a day's neural networks with multiple hidden layers are popular and play a vital role in several huge computational purposes. One great example of neural networks with multiple hidden layers is the Multi-Layer Preceptor (MLP). An MLP can even solve a complex problem which a simple neural network cannot[8].

3.2 Evaluation

3.2.1 Prediction model performance evaluation metrics

To predict the performance of the prediction model we need to use accuracy, specificity, sensitivity, confusion matrix: precision, recall, F1.We calculate true positive(TP), True Negative(TN), false Positive (FP) and false Negative (FN).

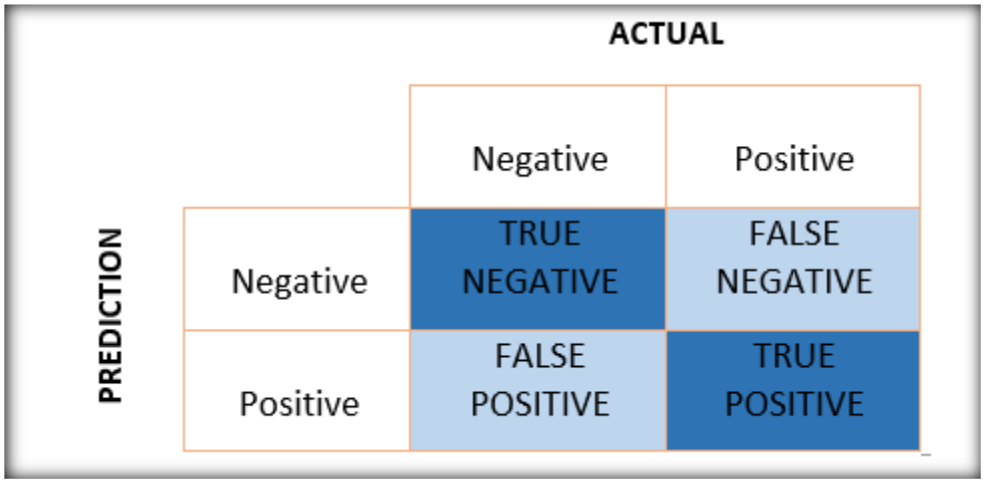


Figure 3.4 Confusion Matrix

True Positive (TP) is a condition that prediction model correctly predicted the positive class which means the prediction and the actual class are positive. True Negative (TN) is a condition that prediction model correctly predict the negative class which means both the prediction class and actual class are negative. False Positive (FP) is a condition that prediction model gives the wrong prediction of the negative class which means prediction class is positive and actual class is negative. False Negative (FN) is a condition that the prediction model wrongly predict the positive class which means prediction class is negative and actual class is positive

3.2.1.1 Accuracy

Accuracy means the algorithm capability to predict the class of the dataset correctly. It determine of how close or near the predicted value is to the actual value[30][41]. Accuracy computes using the following equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad \text{Equation 3. 2}$$

3.2.1.2 Precision

Precision measure the true values correctly predicted from the total predicted values in the actual class[42]. Precision quantifies the ability of the classifiers to not label a **negative** example as **positive.** The equation to calculate precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Equation 3. 3}$$

3.2.1.3 Recall

Recall determine the rate of the total positive values that are correctly predicted. Recall answers what percentage of actual Positives is correctly classified[43].The equation to calculate Recall calculated is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Equation 3. 4}$$

While the macro average is used to calculate the recall value of the prediction models, the equation to calculate macro average recall is:

$$\text{Recall\_macro} = \frac{1}{|\text{C}|}\sum_{\text{i}=1}^{|\text{C}|}\frac{\text{TPi}}{\text{TPi} + \text{FNi}} \qquad \text{Equation 3. 5}$$

Where: $\text{C} = \text{class}$

3.2.1.4 F-measure

F-measure is also called F1-score is the harmonic mean between recall and precision. It takes both false positive value and false negatives values into account. The equation of the F1score is:

$$\mathrm{F1}_{\mathrm{score}} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad \text{Equation 3. 6}$$

The macro average of F1_score is calculated as:

$$\mathrm{F1}_{\mathrm{score\_macro}} = 2 * \frac{\mathrm{Precision\_macro} * \mathrm{Recall\_macro}}{\mathrm{Precision\_macro} + \mathrm{Recall\_macro}} \qquad \text{Equation 3. 7}$$

3.2.1.5 Sensitivity

Sensitivity (SN) is as wellas True Positive Rate (TPR). Sensitivity is the mean proportion of actual true positives that are correctly identified[44].

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{Equation 3. 8}$$

3.2.1.6 Specificity

Specificity (SP) is as well as True Negative Rate (TNR), is calculated the fraction of negative values that are correctly classified.

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad \text{Equation 3. 9}$$

CHAPTER FOUR

SYSTEM MODELING AND EXPERMENTAL SETUP

This part of our study discusses the experimental setup of our proposed prediction model using machine learning algorithm. We establish the classifier problem using different Machine learning supervised algorithms; these are Decision Tree classifier, Random Forest Classifier, Support Vector Machine and Artificial Neural Network. We first provide a detailed description of the experiments and statistical analysis of the dataset and how the experiment was carried out. Finally, a description of how each model achieves the best possible performance was carried out is provided

4.1 Data Description

Our dataset was collected from local hospital that is tikure anbesa hospital. The dataset was new because in Ethiopia there is no published data set for chronic kidney disease prediction model. Tikure anbesa specialized hospital data was organized using excel file but somehow it needs to prepare in the form of the machine learning model. Then after we prepared the dataset, we perform Experimental set up for training and testing the dataset and try to train the model using Decision Tree classifier, Random Forest Classifier, Support Vector Machine, Artificial Neural Network algorithms to predict whether the patient have CKD or Not CKD. After that we evaluate the model accuracy and performance of each model and select the better one that fit for our problem.

The data for this study was collected from Tikur Anbesa Specialized Hospital, and then we have prepared the dataset for prediction based on the methodology discussed in chapter three. The dataset is containing2135 records and 20 features and one class. This dataset is used for classification of either CKD or non-CKD.Table3 presents a list of the attributes, their types and values. Our dataset contains socio-demographic features which is Age and Gender, electrolyte features that include sodium, potassium, and chloride, laboratory test and urine analysis include serum creatinine, albumin, blood urine nitrogen, hemoglobin, platelet count, mean cell volume, red blood cell count, white blood cell count, blood pressure, risk factors features such as hypertension, diabetic mellitus,

anemia, heart disease, and dependent class (Classification). Our dataset is different by some features from online dataset that found in Machine Learning Repository. The online CKD dataset collected from university of California at 2015. In our dataset include Gender, chloride, mean cell volume, platelet count but these are not including UCI CKD dataset. Some of features include in UCI CKD dataset such as sugar, Red blood cell count, Appetite, pedal edema is not including in our dataset. The prepared dataset had missing values and categorical data that need to be handled for building predictive models.

4.2 Data Preprocessing

4.2.1 Identify and handling missing values

The dataset contains missing value and that was identified using isnull() function in python code. The result is showed on figure 4 and discuss in section 3.2.1.4. In our dataset there is 15 numerical features in different range of values, 5 nominal variables which their value defined as “yes” or “No” and one target feature “CKD” or “NOTCKD”. The categorical features contain no missing values but the numerical value has missing values.

The following figure shows the missing value in our dataset, the black color describe the dataset is perfect and no missing values and the white color shows the missing value exist in the respective feature values

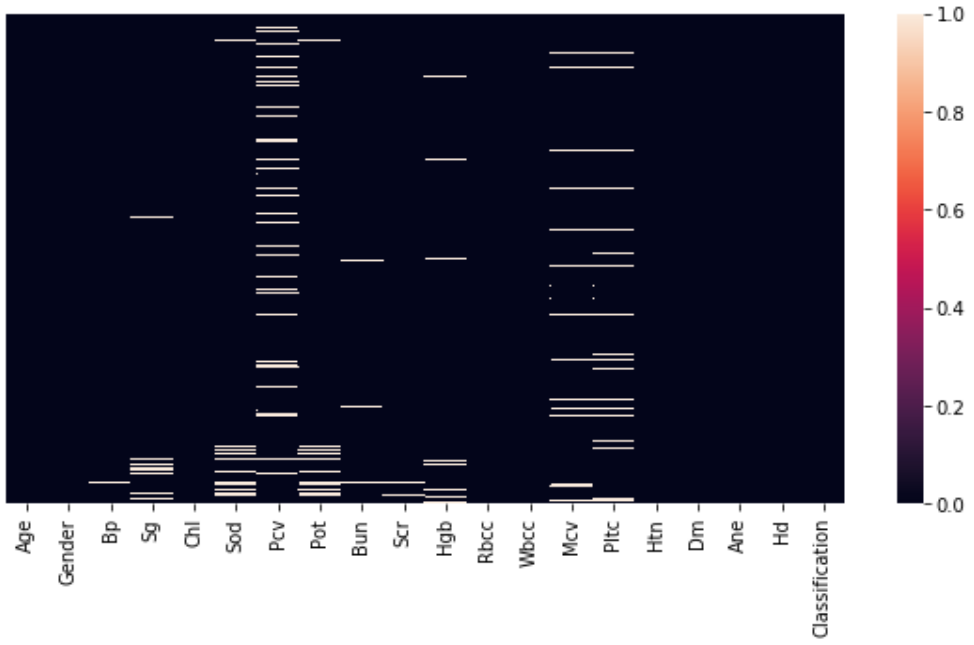


Figure 4.1 Detection of missing values

The missing value were replaced using simple Imputer method, this method replace missing values by using the strategy of mean, median, mode, most frequent and constant for numerical and categorical features. In our dataset case we impute the missing value using mean method.

4.2.2 Data transformation and Encoding

our dataset has categorical features such as Gender, Htn, Ane, HD have nominal value and target feature also have nominal value which is notCkd and ckd that discusses in section 3.2.1.3. All this categorical features have string values and these features are converted into numerical values in the form of 0 and 1 using LableEncoder method form sklearn libraries. Now all dataset features are converted into float64 data type.

4.2.3 Feature Scaling

As we discuss in section 3.2.1.5 feature scaling which it basically helps to normalize the data within a particular range and it also helps in speeding up the calculations in an algorithm. We use min max standardization on numerical features.

4.2.4 Feature Selection

There are several feature selection methods that we discuss in section 3.1.2 in this study we use Pearson correlation method for feature selection in our data preprocessing. Pearson correlation more precisely it is a measure of the extent to which two variables are related [45]. In our feature selection correlation analysis is performed which is if independent feature is highly correlated with dependent feature we need not remove that feature because that kinds of features are play an important role when train our machine learning model and in prediction analysis. But if we have independent features highly correlated each other by more than 90% that means they are duplicated features then we have to remove one of them from our dataset.

4.3 Proposed Chronic Kidney Disease Prediction Model Experimental Setup

The experimental setup was performed to predict and compare the accuracy of four machine learning techniques, i.e. support vector machine, Decision tree, Random forest and ANN, using CKD dataset. The machine learning algorithms were performed on the dataset using python packages. Before that we perform data preprocessing on the dataset, select appropriate features and apply techniques to partition the dataset for training and testing.

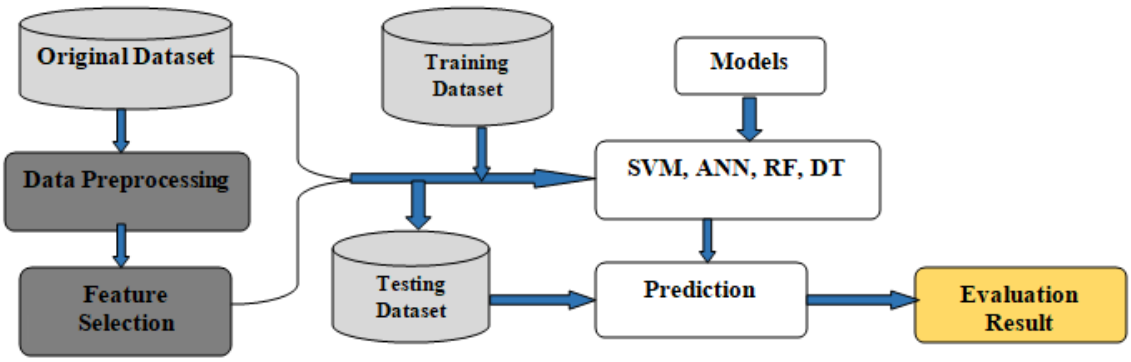


Figure 4.3: CKD prediction Model Architecture

Figure 4.3 shows the overall architecture of our CKD prediction model. This study uses four machine learning algorithm to develop our prediction model

The first proposed classifiers algorithm was support vector machine. Support vector machine have two types, each type used for different purpose. The simple SVM used for linear regression and classification problems and the second one is kernel SVM, it is more flexible for non-linear data because it can add more features to fit a hyper plane instead of two dimensional spaces. We use Kernel SVM for our classification problems, In Kernel function there are different types, i.e. linear, polynomial, Radial Basis functions(RBF), sigmoid. We use these types of function and select the better kernel SVM function.

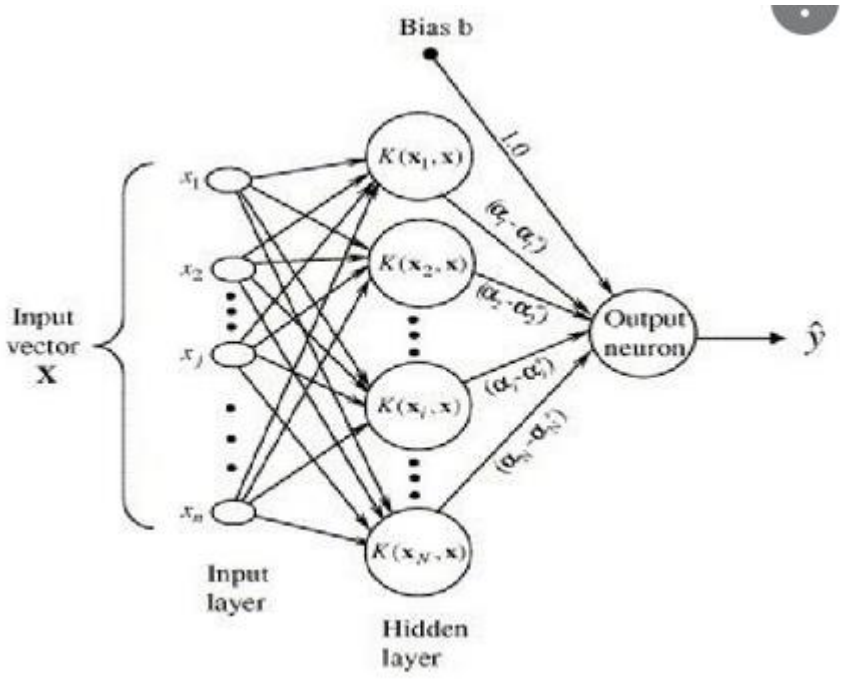


Figure 4.4: Architecture of SVM [46]

The second Algorithm was Random Forest Algorithm. Random forest builds a lot of decision trees and merge all of them together to get more accurate prediction. This model used two concepts: these are random sampling of training data points when building trees and second random subsets of features considered when splitting nodes.

The third classifier algorithm was Decision Tree algorithm. Decision tree learning algorithm can be used for classification and regression problems, in our case uses decision tree classification to predict CKD. DT is made of Root, nodes, branches ad leaves

The fourth classifier Algorithm was Artificial Neural Network. Artificial Neural Network (ANN) is the techniques used in this study to predict chronic kidney Disease. We use Multi-layer preceptor types of artificial neural network. MLP is a category of feed forward artificial neural network that composed of multiple layer of perceptron with activation function. A MLP is consists a minimum of three layers of nodes.

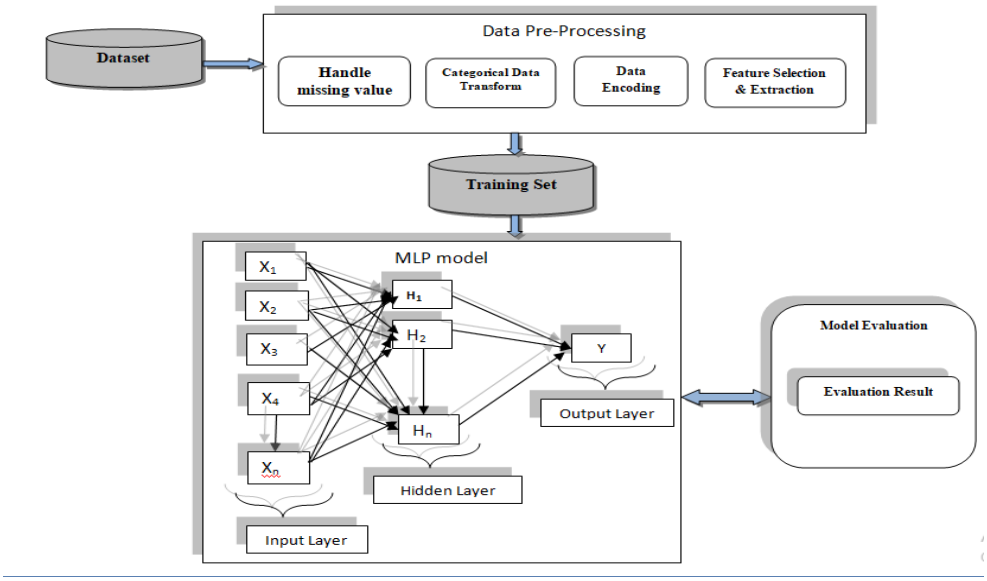


Figure 4.5: Architecture of MLP Model

A multilayer perception is preferred which solves non-linear problems. It is a class of Artificial Neural network. As mention earlier, the algorithm has an input layer, hidden layers, and an output layer. Hidden layers can have any number of nodes but not more than input neurons. By varying the number of nodes in the hidden layer, the accuracy can be varied. Also varied different parameters such as the learning rate, number of neurons, and the number of training iteration by varying the accuracy is varied, the learning rate

should be minimized because the higher learning rate makes the system complex. For each iteration, Accuracy, recall, precision, F1-Score, sensitivity and specificity are calculated and using these methods we evaluate our prediction model

CHAPTER FIVE

IMPLEMENTATION AND DISCUSSION

This chapter explains our chronic kidney prediction model implementation using the defined experimental setup in chapter four

5.1 Implementation of prediction model Environment

In this part of our study we discuss the working environment of the proposed prediction model and we set up the environment according to our research needed. We use the python programming language to implement the experimental setup that we discuss in chapter 4. Python programming is a kind of general-purpose programming language that we can used for several things like web development, AI, machine learning, operating systems, mobile application development, and video games. To install the python package we use Anaconda navigation version 4.5.11 for the distribution of python packages. For this study we choose python because of easy to understand and allows for quick data validation, easy access the several libraries like pandas, sklearn, Matplotlib …etc, testing can be performed easily because tests can be run on any platform and several researchers uses Python for develop machine learning model. The following table shows all python libraries and packages that we used in this study

Table 5.1. Description of the Tools and Python Packages Used During the Implementation

Tools and packages	Version	Description
Anaconda navigator	4.5.11	Used for the distribution of python packages that have many libraries, working environments Specially for building the machine learning model.
Jupiter Notebook	3.7.0	It is a free, open-source, interactive web tool which known as a computational notebook that enable us to usesoftware code, computational output, explanatory text and several recourses for Data preprocessing and model development.
Python	3.7.0	Python is a popular platform easy to learn, powerful programming language to develop a machine learning

		application.
Microsoft Excel	2010	To prepare the dataset
Microsoft Word	2010	To prepare research documentation
Sickie-learn	0.21.3	It is a python module that we uses for developing our classification models
Pandas	0.23.4	Pandas are an open-source python library that allows us to perform data manipulation and analysis for numerical data, loading dataset from Excel files and handling of dataset.
Numpy	1.19.5	It is an open source python module that we used for handling categorical data in to numerical data, multi-dimensional array processing and objects.
Seaborrn	3.1.1	Statistical data visualization
Matplotlib	0.9.0	It is a graph plotting library in Python to we used as visualization of data distribution, and several results

5.2 Deployment Environment

In this study the above table (Table 5.1.) discussed tools and modules are implemented and install on personal computer properties are Intel(R) Core(TM) i3-6006U CPU @ 2.00GHz, 2000 Mhz, 2 Core(s), 4 Logical Processor(s), 4 GB RAM, 500 HD storage and Microsoft Windows 10 Pro.

5.3 Data Preprocessing Implementation

This study uses a python programming for Data preprocessing on the chronic kidney dataset. The tasks of all data preprocessing such as handling missing values, converting categorical, feature selection and feature scaling. All nominal and categorical data should be converted to 0 and 1, to handle missing value we used Mean strategy. First we import all libraries to Jupiter notebook.

```
#import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from pandas import Series, DataFrame
import missingno as msno
from sklearn.impute import SimpleImputer
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.neural_network import MLPRegressor
from sklearn.neural_network import MLPClassifier
```

Figure 5.1Python code for the import python library

The above figure shows all the python code Libraries and packages that we used in our chronic kidney prediction model. All of the libraries are described in table 4.

```
#Load the Dataset
data=pd.read_csv('Prediction_CKD.csv')
#Print the first Five Rows
data.head()
```

Figure 5.2Python code for loading the dataset.

5.3.1 Implementation of handling missing values in the dataset

In order to handle the missing values in the in our dataset, first we have to identify the missing value using Isnull() method and calculate the total amount of missing value for each futures using sum() methods used the SimpleImputer() method that replaces the missing values by calculating the mean value.

```
#handle missing values using simple imputer method
from sklearn.impute import SimpleImputer
imp=SimpleImputer(missing_values=np.nan,strategy='mean')
data['Bp']=imp.fit_transform(data[['Bp']])
data['Ane']=imp.fit_transform(data[['Ane']])
data['Sg']=imp.fit_transform(data[['Sg']])
data['Chl']=imp.fit_transform(data[['Chl']])
data['Sod']=imp.fit_transform(data[['Sod']])
data['Pot']=imp.fit_transform(data[['Pot']])
data['Bun']=imp.fit_transform(data[['Bun']])
data['Scr']=imp.fit_transform(data[['Scr']])
data['Hgb']=imp.fit_transform(data[['Hgb']])
data['Rbcc']=imp.fit_transform(data[['Rbcc']])
data['Wbcc']=imp.fit_transform(data[['Wbcc']])
data['Mcv']=imp.fit_transform(data[['Mcv']])
data['Pltc']=imp.fit_transform(data[['Pltc']])
data['Pcv']=imp.fit_transform(data[['Pcv']])
```

Figure 5.3 Python code for handling missing values

5.3.2 Data transformation and Encoding

All categorical features have string values and these features are converted into numerical values in the form of 0 and 1 using LableEncoder method form sklearn libraries.

```
#Transform the non-numeric data in the column
for columns in data.columns:
    if data[columns].dtype == np.number:
        continue
    data[columns]=LabelEncoder().fit_transform(data[columns])
```

Figure 5.4 Python code for Data Transformation for categorical features

```
#split the data into 80% training and 20% testing
X_train,X_test,Y_train,Y_test=train_test_split(X,Y, test_size=0.3, shuffle=True)
```

Figure 5.5 python code for Split Dataset

```
#feature scaling
#min-max scalar method scaled the data set so that all the input features between 0 and 1
x_scalar=MinMaxScaler()
x_scalar.fit(X)
columns_name=X.columns
X[columns_name]=x_scalar.transform(X)
```

Figure 5.6Python code for feature scaling

5.4 Implementation of Feature Selection

First we split our dataset independent(X) dataset (the feature) and dependent(Y) dataset (the target).Then we perform feature scaling and Pearson correlation analysis to perform feature selection.

```
#with the following function we select highly correalted features
# it will remove the first feature that is correlated with any thing other feature
def correlation(dataset, threshold):
    col_corr=set() #set of all the ames of correlated columns
    corr_matrix=dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if abs(corr_matrix.iloc[i,j])>threshold: #we are interested in absolute coeff value
                colname=corr_matrix.columns[i] # select the name of columens
                col_corr.add(colname)
    return col_corr
```

```
#show higly correlated features
corr_features=correlation(X_train, 0.7)
len(set(corr_features))
```

```
0
```

```
corr_features
```

```
set()
```

```
# drop higly correlated features from the dataset
X_train.drop(corr_features, axis=1)
X_test.drop(corr_features, axis=1)
```

Figure 5.8. Pearson correlation python code

5.5 Machine Learning Models Implementation

In this study we used to build a machine learning model with python Ski-learn library and as we discussed in chapter 4 the CKD prediction model experimental set up with four Machine learning models and implement them separately with their specific parameters.

The first classifier algorithm we used in this study is support vector machine. The study used the kernel SVM type of functions called SVC() classifier of the sklearn package to building the SVM model. This classifier is initializing with a basic parameter like Kernel parameter value such as RBF and Linear, polynomial and sigmoid. After computing the accuracy of all types of kernel parameters then we choose linear with better accuracy to classify the prediction model..

```
#SVM classifier
from sklearn import svm
svmclas=svm.SVC(kernel='linear', gamma='auto',C=40,random_state=1)
svmclas.fit(X_train,Y_train)
```

Figure 5.9. Python code for SVM Model

The second classifier we used in our research is Decision Tree classifier with several parameters. The ski-learn library method for decision tree classifier is DecisionTreeClassifier() function.

```
#Decision Tree
from sklearn.tree import DecisionTreeClassifier
tree=DecisionTreeClassifier(criterion='entropy', random_state=0)
tree.fit(X_train,Y_train)
```

Figure 5.10. python code for DTModel

RF model is implemented through Random Forest Classifier () method of the sklearn ensemble package is used to train the classifier. This classifier fits several decision tree classifiers as declare in n_estimators parameters.

```
#Random Forest Classification
from sklearn.ensemble import RandomForestClassifier
forest=RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=0)
forest.fit(X_train,Y_train)
```

Figure 5.10. python code for RF Model

Multi-layer Perseptron Artificial neural network model is implemented through MLPClassifier() method of the sklearn ensemble package is used to train the classifier. This classifier uses several attributes like activation, Learning rate, .etc

```
#MLP model with MLPClasifier
from sklearn.neural_network import MLPClassifier
MLP=MLPClassifier(hidden_layer_sizes=(20,80,30,10,5), activation='relu',verbose=True, max_iter=8000,
                  solver='adam', random_state=30,alpha=0.005, learning_rate_init=0.005)
MLP.fit(X_train,Y_train)

print()
print('[1] Multi-Layer Classifier Trainig Accuracy :',MLP.score(X_train,Y_train))

print()
print()
#Evaluuate MLP Classifier
print('[1] Multi-Layer Classifier Testing Accuracy :')
from sklearn.metrics import classification_report, accuracy_score
print(classification_report(Y_test,MLP.predict(X_test)))

print()
```

Figure 5.11. Python code for Multi-layer artificial neural network Model

5.6 Model Testing and Evaluation

In this part of our study we used to perform model testing and evaluation of the chronic kidney disease prediction model using performance evaluation metrics including precisions, recall, recall and fl-score.

```
#confusion matrix of all models
from sklearn.metrics import classification_report, accuracy_score
for i in range(len(model)):
    print('model ', i)
    print(classification_report(Y_test,model[i].predict(X_test)))
    print(accuracy_score(Y_test,model[i].predict(X_test)))
    print()
```

Figure 5.12. Python code for Model evaluation

5.7 Discussion

In this point we discuss the outcome that we implement the experimental setup using four machine learning algorithm to develop our proposed chronic kidney disease. At this section our dataset class distribution discussed and each machine learning prediction model result are discussed.

5.7.1 Dataset collection and Preprocessing Result

The collected dataset was total of 2135 and after preprocessing still the dataset have a record of 2135This means that we cannot drop any data record from our dataset and we handle the dataset missing value properly using simpleimputation() method. Out of the total dataset, 1700 patients have a Chronic Kidney disease and 435 have no chronic kidney disease or normal. The class distribution of our Dataset is shown in the following figures 5.13.

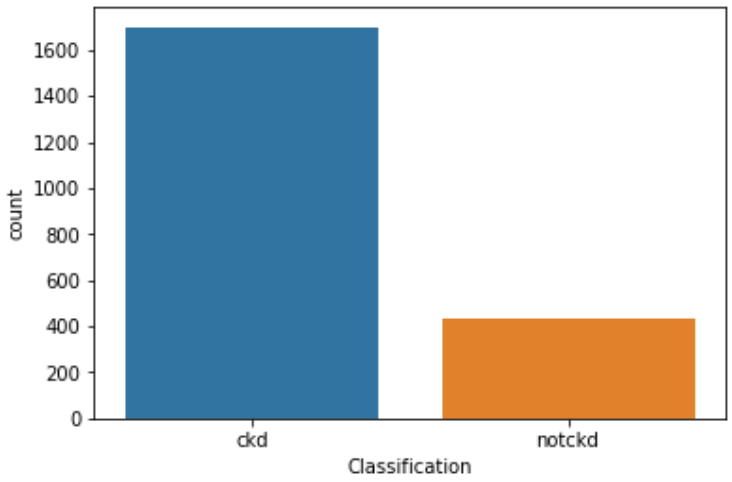


Figure 5.13. Dependent (Classification) feature value distribution

The above figure shows in our dataset have imbalance class distribution. An imbalanced dataset means instances of one of the two classes is higher than the other. Most of the classification algorithms get impacted by imbalanced datasets. The problem with training the model with an imbalanced dataset is that the model will be biased towards the majority class only[47]. This causes a problem when we are interested in the prediction

of the minority class. To solve this problem either of two methods are used the first we need to convert imbalanced datasets into balanced datasets or second use evaluation 'Precision/'Recall' or F1-score performance measure for imbalanced datasets[48].Accuracy is considering a good evaluation when we have balance dataset. Because when we have the balance dataset the prediction cannot biased with one of the higher categories. Because of that Accuracy evaluation gives false results for imbalanced datasets

In this study to handle this imbalance problem we use two methods. The first option is using F1-score, precision and recall evaluation. In this way our prediction models using confusion matrix to show the performance of each models.

In the second method we use to convert imbalanced datasets into balanced. To convert the imbalance dataset in to balance dataset there are two methods are presented from several researchers. These are, Over Sampling which means converting minority class number is equal to majority class number and Under Sampling means converting the majority class number is equal to minority class number. We focus on the first method over sampling because the second method in our dataset case simply removes samples from the majority class, with or without replacement. This is one of the earliest techniques used to alleviate the imbalance in the dataset, but, it may increase the variance of the classifier and may potentially discard useful or important information from our dataset. This leads to bias for the prediction model. The oversampling technique was used to balance the value of the minority class with the value of the majority class. After the resembling method was used, the size of class labels for Classification class attributes was balanced to 1700 ckd and 1700 notckd and total of 3400 as shown in the following figure.

```
from imblearn import under_sampling, over_sampling
from imblearn over_sampling import RandomOverSampler
ros=RandomOverSampler(random_state=0)
X_resampled,Y_resampled=ros.fit_resample(x,y)
print(sorted(Counter(Y_resampled).item(),Y_resampled.shape())
```

Figure 5.14. sample code for resample imbalance dataset

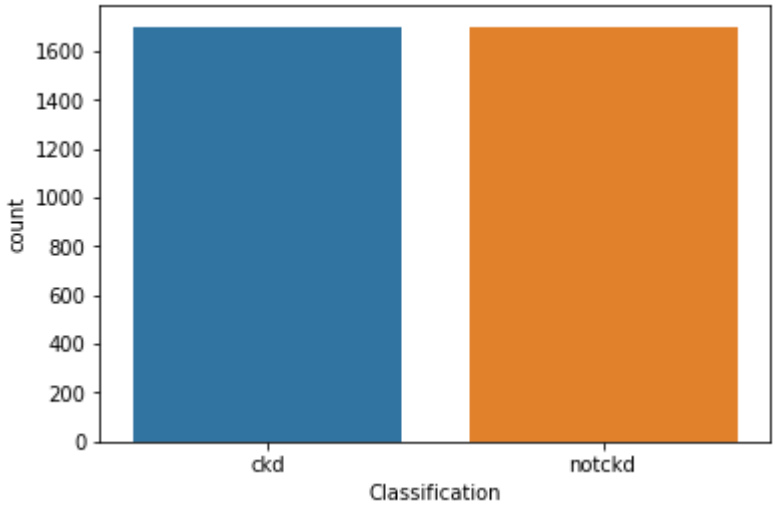


Figure 5.15. Class distribution after recompleting

5.1.1 Feature selection Result

The feature selection process in this study uses Pearson correlation analysis method. As a result of feature selection using this method we cannot remove any feature from our dataset because in the correlation analysis function, we define a threshold value which is greater than80% correlated one feature from the other feature, these two features supposed to having the same values then we have to remove one of the two features. But in our case there is no feature correlation to satisfy that threshold parameter. The highest correlation having between serum creatinine and blood urea nitrogen with 66% correlation, hemoglobin with red blood cell count also have 66% correlation each other. We also consider negative correlation, negative correlation is considered in this study by defining absolute function because negative correlation is also very important which indicates a feature increase in its value and the other value is decrease in other side so we

should not remove this kinds of correlation because it gives information about their association with different direction, and that may useful for machine learning algorithm. Allresult shown in figure 5.16

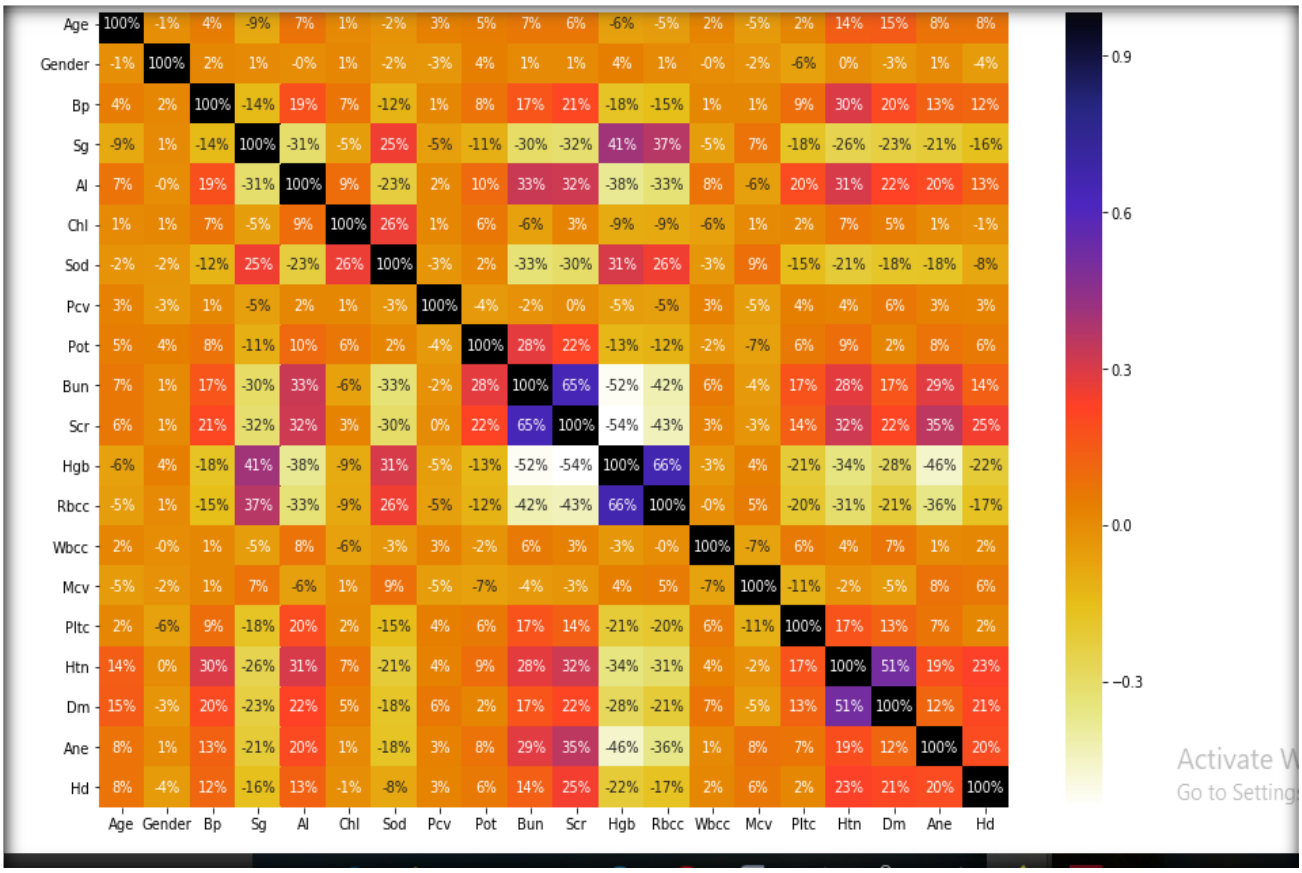


Figure 5.16. Correlation of independent variables

5.1.2 Model Evaluation Results

In Our prediction model the Experiment set up was using 4 machine learning algorithms which are DT, RF, SVM and ANN. First we train the model using different parameters and evaluate using performance evaluation matrix. the result of each model were discussed in this section.

First we train our model using Decision tree and Random forest algorithms and the accuracy we get 95.9% and 96.2% respectively, Recall (Weighted Avg), F-Measures, precision are 99%. The following table5.1 shows the evaluation matrix result of Decision tree and Random forest model.

Table 5.1Outcome of Decision tree and Random Forest Algorithm

Evaluation Matrix	**Decision tree Model result**	**Random Forest model result**
Total number of instance	681	681
Accuracy	95.9%	96.2%
Recall (Weighted Avg)	0.96	0.96
F1-Scorre (Weighted Avg)	0.96	0.96
Precision (Weighted Avg)	0.96	0.96

Secondly we train our model using support vector machine algorithms using different parameters andthe best result was found on kernel type linear and C value =10 the accuracy we get 98.2%, Recall (Weighted Avg), F-Measures, precision are 98%

Table 5.2Outcome of SVM Algorithm

	Kernel type	**C**	**Evaluation Matrix**				
			Total number of instance	Accuracy	Recall (Weighted Avg)	F1-Scorre (Weighted Avg)	Precision (Weighted Avg)
1	linear	40	681	98.2%	0.98	0.98	0.98
2	rbf	40	681	95.8%	0.96	0.96	0.96
3	rbf	10	681	94.4%	0.94	0.94	0.94
4	linear	10	681	98.9%	0.98	0.98	0.98

Table 5.3Outcome of MLP ANN Algorithm

#neurons	# Hidden layers	Activation	Solver	Evaluation Matrix				
				Total number of instance	Accuracy	Recall (Weighted Avg)	F1-Scorre (Weighted Avg)	Precision (Weighted Avg)
20,80,30,10,5	5	relu	adam	681	98.9%	0.99	0.99	0.99
20,100,50,10	6	relu	adam	681	97.7%	0.98	0.98	0.98
20,100,50, 10	4	relu	lbfgs	681	98.2%	0.99	0.99	0.99

Table 5.3 shows Multilayer perceptron classification model with the number of nodes in the hidden layer, activation, solver and maximum iteration number. By varying this attributes for our classification model we get different result. This result evaluated by performance evaluation methods and confusion metrics.

Choosing the best activation function and number of hidden layers If the number of hidden layers gets increased or decreased and for different activation function, the accuracy of testing and training data relatively equal, but the activation function plays a major role in getting higher accuracy with small amount of computational time. Some of the Comparison for different activation function, hidden layer, number of neurons and maximum iteration with confusion matrix evaluation for our model are represented in the table5.3.

By comparing different parameters for MLP algorithm we get best result on number of hidden layer= 5, Activation = relu, Solver=adam and number of neurons at each hidden layer respectively = 20, 80, 30,10,5. With these parameters the accuracy we get 99.0%, Recall (Weighted Avg), F-Measures, precision are 99%

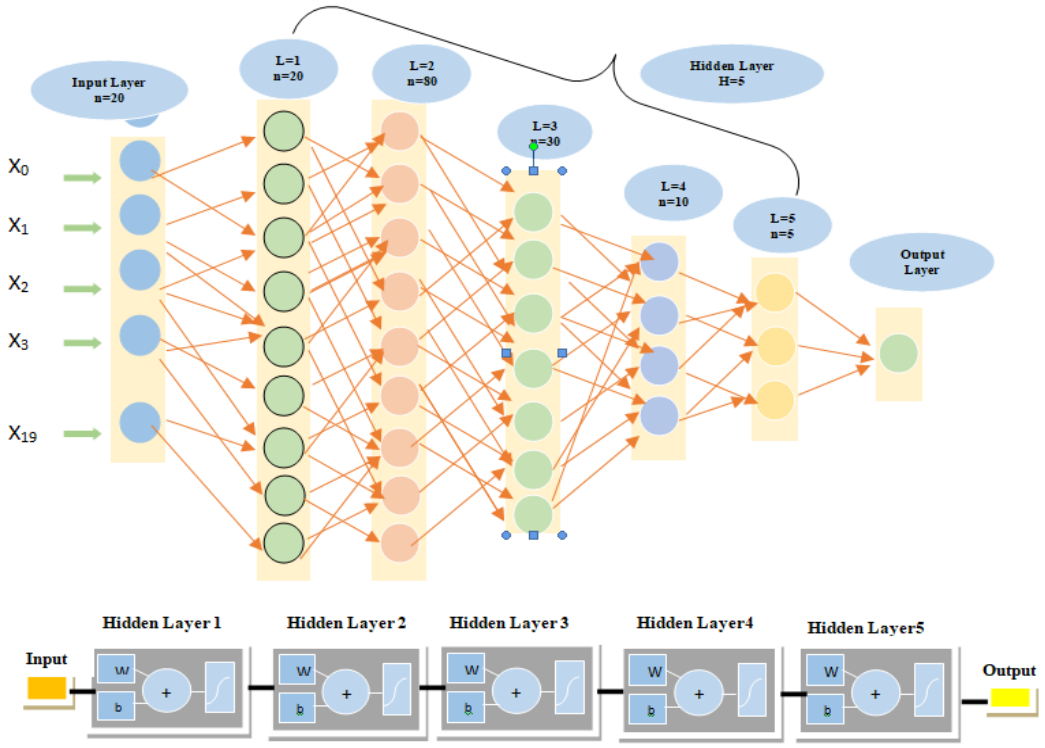


Figure 5.17. Architecture of MLP ANN

```
[1] Multi-Layer Classifier Trainig Accuracy : 0.9856617647058824


[1] Multi-Layer Classifier Testing Accuracy :
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       332
           1       1.00      0.98      0.99       349

    accuracy                           0.99       681
   macro avg       0.99      0.99      0.99       681
weighted avg       0.99      0.99      0.99       681
```

Figure 5.17. Evaluation result for MLP ANN

Finally, we summarize all algorithm result in the following table 5.4

Table 5.4Result summary of all Algorithms

Evaluation Matrix	Decision tree Model result	Random Forest model result	SVM Model result	MLP Model result
Total number of instance	681	681	681	681
Accuracy	95.5%	96.2.2%	98.9%	99.0%
Recall (Weighted Avg)	0.96	0.96	0.98	0.99
F1-Scorre (Weighted Avg)	0.96	0.96	0.98	0.99
Precision (Weighted Avg)	0.96	0.96	0.98	0.99

Decision trees are much easier to train and understand the classification model. Since a random forest combines multiple decision trees, it becomes more difficult to train and understand the classification model relatively to Decision tree specially when the number of record increase in the dataset Random Forest has a higher training time than a single decision tree. we should take this into consideration because as we increase the number of trees in a random forest, the time taken to train each of them also increases. Support vector machine also have a good result but relatively low accuracy that of MLP ANN and it doesn't perform well if the dataset increased. As we discuss in table 2.1 neural network also is good to model with nonlinear data with large number of inputs and once trained, the predictions are pretty fast and can be trained with any number of inputs and layers which is best with more data points but it also need computational resources.

In this study we get relatively the same result from all prediction models which is accuracy The results get from DT was 95.5%, SVM was 98%, RF was 96.5% and ANN get accuracy 99%, relatively the same accuracy

CHAPTER SIX

CONCLUSION AND FUTURE WORK

In this section, the proposed chronic kidney disease prediction system using machine-learning techniques concluded and future research directions have been described.

6.1 Conclusion

Chronic kidney disease (CKD) is a worldwide health problem that affects people from their day to day activities. Especially it is a major challenge in low income countries like Ethiopia so that this prediction model helps the patient and the health care system to easily identify the CKD problem and slow down the progression of the CKD problem by maintaining the patient day to day life style. This research, chronic kidney disease prediction model using machine learning algorithm is proposed. Data preprocessing is performed to make the data clean and make it suitable for machine learning models.

In this study, the missing values was handled using the simple imputation method and categorical values have been converted to 0 and 1. Therefore, there was a class imbalance in the dataset. The over sample technique was used to make balance the dataset class distribution. The cleaned total dataset size is 3401. The model is developed using this clean dataset. We use DT, SVM, RF and Multi-Layer Preceptor Artificial Neural Network machine-learning models are used to develop Prediction model and Pearson correlation feature selection methods were used to build the proposed model. The evaluation of the model was done using confusion matrix with precision, recall and F1-score. Applying the models on the dataset, we have got the highest accuracy. The results of this study will get from DT was 95.5%, SVM was 98.2%, RF was 96.5% and ANN get accuracy 99%. relatively from each model but this study recommends ANN model for better prediction for future because when the number of features and records increase it gives best result but it needs more computational resource.

6.2 . Recommendation

Currently CKD become an international problem and many people die from this disease in Ethiopia even if they don't know that had chronic kidney because CKD cannot show any symptoms until it reaches ESRD. The applications of machine learning models are used in health care system by providing decision support for the expertise.

In our country health provision centers and hospitals have information technology infrastructure in their hospitals and service centers but their data management is weak after face a huge challenge of data collection in Tikure Anbesa specialized hospital, we observe most hospitals uses their computers to record laboratory results and reporting purpose rather than using their computers for decision support system, there is no integration of health care system. Tikur Anbesa establish EHIS before two years but the system is not fully integrated and the expert cannot use a technology, still every lab results record on papers and analyze by the doctors so that sometimes this is so difficult. If laboratory result directly records on computer and analyze by this type of applications, then the problems are easily identified and make decisions.

6.3 Future Works

This study used a DT, SVM, RF and Multi-Layer Artificial neural network algorithm to develop the prediction model for chronic kidney disease. The dataset was imbalanced for classification problem that we used by this model. The dataset organization in local hospitals is poor and it is not well organized. Therefore, it is better to model with balanced data to achieve the highest accuracy of the model and in future it needs more work on determines the magnitude of CKD among high blood pressure, hypertension and diabetics disease and examine their relationship between these disease and CKD. This help to reduce the occurrence of CKD

REFERENCES

[1] H. Kriplani, B. Patel, and S. Roy, *Prediction of chronic kidney diseases using deep artificial neural network technique*, vol. 31. Springer International Publishing, 2019.

[2] S. Y. Yashfi *et al.*, "Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225548.

[3] Y. Tadesse and S. Lulseged, "Editorial Kidney Transplantation in Ethiopia : the Dawn of a New Era in Renal Care," pp. 1–2, 2020.

[4] H. Alemu, W. Hailu, and A. Adane, "Prevalence of Chronic Kidney Disease and Associated Factors among Patients with Diabetes in Northwest Ethiopia: A Hospital-Based Cross-Sectional Study," *Curr. Ther. Res. - Clin. Exp.*, vol. 92, p. 100578, 2020, doi: 10.1016/j.curtheres.2020.100578.

[5] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[6] P. Foëx and J. W. Sear, "Hypertension: Pathophysiology and treatment," *Contin. Educ. Anaesthesia, Crit. Care Pain*, vol. 4, no. 3, pp. 71–75, 2004, doi: 10.1093/bjaceaccp/mkh020.

[7] Z. Wang, J. Won Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique," *Int. J. Eng. Technol.*, vol. 7, no. 4.36, p. 1161, 2018, doi: 10.14419/ijet.v7i4.36.25377.

[8] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, no. October 2018, pp. 101–111, 2019, doi:

10.1016/j.compbiomed.2019.04.017.

[9] G. Chen *et al.*, “Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform,” *IEEE Access*, vol. 8, pp. 100497–100508, 2020, doi: 10.1109/ACCESS.2020.2995310.

[10] A. Pinto, D. Ferreira, C. Neto, A. Abelha, and J. Machado, “Data mining to predict early stage chronic kidney disease,” *Procedia Comput. Sci.*, vol. 177, no. 2018, pp. 562–567, 2020, doi: 10.1016/j.procs.2020.10.079.

[11] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, “Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 25407–25419, 2020, doi: 10.1109/ACCESS.2020.2971208.

[12] N. V. Ganapathi Raju, K. Prasanna Lakshmi, K. G. Praharshitha, and C. Likhitha, “Prediction of chronic kidney disease (CKD) using Data Science,” *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 642–647, 2019, doi: 10.1109/ICCS45141.2019.9065309.

[13] A. L. Barbieri, O. Fadare, L. Fan, H. Singh, and V. Parkash, “Challenges in communication from referring clinicians to pathologists in the electronic health record era,” *J. Pathol. Inform.*, vol. 9, no. 1, pp. 1–5, 2018, doi: 10.4103/jpi.jpi.

[14] Q. Li *et al.*, “Machine learning in nephrology: scratching the surface,” *Chin. Med. J. (Engl).*, vol. 0, no. 6, pp. 687–698, 2020, doi: 10.1097/CM9.0000000000000694.

[15] P. Kotturu, V. V. S. Sasank, G. Supriya, C. S. Manoj, and M. V. Maheshwarredy, “Prediction of chronic kidney disease using machine learning techniques,” *Int. J. Adv. Sci. Technol.*, vol. 28, no. 16, pp. 1436–1443, 2019, doi: 10.17148/IJARCCE.2018.71021.

[16] K. C and Y. v, “Prevalence of Chronic Kidney Disease and Associated factors among Patients with Kidney Problems Public Hospitals in Addis Ababa, Ethiopia,” *J. Kidney*, vol. 04, no. 01, pp. 1–5, 2018, doi: 10.4172/2472-1220.1000162.

[17] A. Khamparia, G. Saini, B. Pandey, S. Tiwari, D. Gupta, and A. Khanna, “KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network,” *Multimed. Tools Appl.*, vol. 79, no. 47–48, pp. 35425–35440, 2020, doi: 10.1007/s11042-019-07839-z.

[18] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[19] S. W. A. Sherazi, Y. J. Jeong, M. H. Jae, J. W. Bae, and J. Y. Lee, “A machine learning–based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome,” *Health Informatics J.*, vol. 26, no. 2, pp. 1289–1304, 2020, doi: 10.1177/1460458219871780.

[20] L. Yahaya, N. David Oye, and E. Joshua Garba, “A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques,” *Am. J. Artif. Intell.*, vol. 4, no. 1, p. 20, 2020, doi: 10.11648/j.ajai.20200401.12.

[21] F. Ma, T. Sun, L. Liu, and H. Jing, “Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network,” *Futur. Gener. Comput. Syst.*, vol. 111, pp. 17–26, 2020, doi: 10.1016/j.future.2020.04.036.

[22] B. L. Neuen, S. J. Chadban, A. R. Demaio, D. W. Johnson, and V. Perkovic, “Chronic kidney disease and the global NCDs agenda,” *BMJ Glob. Heal.*, vol. 2, no. 2, pp. 1–5, 2017, doi: 10.1136/bmjgh-2017-000380.

[23] T. Shibiru, E. K. Gudina, B. Habte, A. Derbew, and T. Agonafer, “Survival patterns of patients on maintenance hemodialysis for end stage renal disease in Ethiopia: Summary of 91 cases,” *BMC Nephrol.*, vol. 14, no. 1, 2013, doi:

10.1186/1471-2369-14-127.

[24] K. Kumela Goro *et al.*, “Patient Awareness, Prevalence, and Risk Factors of Chronic Kidney Disease among Diabetes Mellitus and Hypertensive Patients at Jimma University Medical Center, Ethiopia,” *Biomed Res. Int.*, vol. 2019, 2019, doi: 10.1155/2019/2383508.

[25] P. T. Coates *et al.*, “KDIGO 2020 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease,” *Kidney Int.*, vol. 98, no. 4, pp. S1–S115, 2020, doi: 10.1016/j.kint.2020.06.019.

[26] L. Chen, “Overview of clinical prediction models,” *Ann. Transl. Med.*, vol. 8, no. 4, pp. 71–71, 2020, doi: 10.21037/atm.2019.11.121.

[27] K. Hempstalk, “Introduction to Machine Learning in Healthcare,” p. 24, 2018, [Online]. Available: http://web.orionhealth.com/rs/981-HEV-035/images/Introduction_to_Machine_Learning.pdf.

[28] T. O. Ayodele, “Atherosclerotic Cardiovascular Disease,” *Atheroscler. Cardiovasc. Dis.*, 2012, doi: 10.5772/711.

[29] A. Abdi, “Three types of Machine Learning Algorithms List of Common Machine Learning Algorithms,” no. November, 2016, doi: 10.13140/RG.2.2.26209.10088.

[30] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, “Applications of machine learning predictive models in the chronic disease diagnosis,” *J. Pers. Med.*, vol. 10, no. 2, 2020, doi: 10.3390/jpm10020021.

[31] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, “Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning,” *IEEE Access*, vol. 7, pp. 152900–152910, 2019, doi: 10.1109/ACCESS.2019.2948430.

[32] A. Al Imran, M. N. Amin, and F. T. Johora, “Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide Deep

Learning," *2018 Int. Conf. Innov. Eng. Technol. ICIET 2018*, pp. 1–6, 2019, doi: 10.1109/CIET.2018.8660844.

[33] H. I. Ozturk, A. Saglik, B. Demir, and A. G. Gungor, "An artificial neural network base prediction model and sensitivity analysis for marshall mix design," no. June, 2017, doi: 10.14311/ee.2016.224.

[34] D. Schmidt, M. Niemann, and G. Lindemann-Von Trzebiatowski, "The handling of missing values in medical domains with respect to pattern mining algorithms," *CEUR Workshop Proc.*, vol. 1492, pp. 147–154, 2015.

[35] S. F. Huang and C. H. Cheng, "A safe-region imputation method for handling medical data with missing values," *Symmetry (Basel).*, vol. 12, no. 11, pp. 1–19, 2020, doi: 10.3390/sym12111792.

[36] A. Palmer, R. Jimenez, and E. Gervill, "Data Mining: Machine Learning and Statistical Techniques," *Knowledge-Oriented Appl. Data Min.*, no. May 2014, 2011, doi: 10.5772/13621.

[37] E. Bahadır, "Using neural network and logistic regression analysis to predict prospective mathematics teachers' academic success upon entering graduate education," *Kuram ve Uygulamada Egit. Bilim.*, vol. 16, no. 3, pp. 943–964, 2016, doi: 10.12738/estp.2016.3.0214.

[38] M. Awad and R. Khanna, "Efficient learning machines: Theories, concepts, and applications for engineers and system designers," *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.*, no. July 2018, pp. 1–248, 2015, doi: 10.1007/978-1-4302-5990-9.

[39] S. Ramya and N. Radha, "Diagnosis of Chronic Kidney Disease Using," pp. 812–820, 2016, doi: 10.15680/IJIRCCE.2016.

[40] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," *IFMBE Proc.*, vol. 62, no. 1, pp. 589–594, 2017, doi: 10.1007/978-981-10-4166-2_89.

[41] A. Acharya, “Comparative Study of Machine Learning Algo- rithms for Heart Disease Prediction,” no. April, 2017.

[42] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[43] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. M. Riad, “A machine learning model for improving healthcare services on cloud computing environment,” *Meas. J. Int. Meas. Confed.*, vol. 119, pp. 117–128, 2018, doi: 10.1016/j.measurement.2018.01.022.

[44] Y. Amirgaliyev, “Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods,” *2018 IEEE 12th Int. Conf. Appl. Inf. Commun. Technol.*, pp. 1–4, 2010.

[45] J. Sim, J. S. Lee, and O. Kwon, “Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications,” *Math. Probl. Eng.*, vol. 2015, 2015, doi: 10.1155/2015/538613.

[46] N. Farid, B. M. Elbagoury, M. Roushdy, and A.-B. M. Salem, “A Comparative Analysis for Support Vector Machines For Stroke Patients,” *WSEAS Proc. 7th Eur. Comput. Conf.*, no. June, pp. 71–76, 2013.

[47] D. Ramyachitra and P. Manikandan, “Imbalanced dataset classification and solutions: a review,” *Int. J. Comput. Bus. Res.*, vol. 5, no. 4, 2014.

[48] Y. Sun, A. K. C. Wong, and M. S. Kamel, “Classification of imbalanced data: A review,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009, doi: 10.1142/S0218001409007326.

APPENDICES

Appendix A: Dataset Features Description

- **Age** is the age of the patient in years. Age is one factor of chronic kidney disease. The prevalence and severity of CKD is higher among patients >60 years old.
- **Gender** is the gender of the patient (Male or Female).
- **Bp**is the blood pressure in mm/Hg. High blood pressure can damage blood vessels in the kidneys, reducing their ability to work properly.
- **Sg**is the specific gravity of the patient (nominal).
- **Chl**is chloride (mmol/L), which is the CKD predictive attribute.
- **Sod** is sodium (mmol/L), which is essential for the human body. A person with CKD cannot eliminate excess sodium and fluid from his or her body. High sodium increases blood pressure.
- **Pot** is potassium (mmol/L), which is important for the human body. A person with a damaged kidney cannot eliminate excess potassium and fluid from his or her body.
- **Bun** Blood Urine Nitrogen (mg/dL) is the CKD predictive attribute.
- **Scr** serum creatinine (mg/dL) is a waste product that comes from muscle activity. Damaged kidneys cannot remove creatinine from the blood as much as the proper kidney. Serum creatinine is the most predictive parameter for CKD detection.
- **Hgb** hemoglobin (gm/dl) is important CKD predictive parameter.
- **Rbcc** red blood cell count (m/cumm).
- **Wbcc** red blood cell count (k/cumm).
- **Htn** hypertension (yes or no) is the factor that increases the progress of CKD.
- **Dm**diabetic mellitus (yes or no) is a good predictive attribute of CKD.
- **Ane** anemia (yes or no) is the factor of CKD.
- **Mcv** Mean cell volume (fl).
- **Pltc** platelet count (k/cumm).
- **Hd**heart disease (yes or no) is the factor of CKD.

- **Classification** the class assigned to the patient data history. (ckd and notckd).

Appendix B: Sample Code

Appendix B.1 Sample Code for Preprocessing

```
#Data transformation used to replace Yes and No value with 0 and 1
data["Htn"]=data["Htn"].map({
   "no":0,
   "yes":1
}.get)
data["Dm"]=data["Dm"].map({
   "no":0,
   "yes":1
}.get)
data["Ane"]=data["Ane"].map({
   "no":0,
   "yes":1
}.get)
data["Hd"]=data["Hd"].map({
   "no":0,
   "yes":1
}.get)
data["Classification"]=data["Classification"].map({
   "notckd":0,
   "ckd":1
}.get)
data["Gender"]=data["Gender"].map({
   "F":0,
   "M":1
}.get)
```

```
#Encode Object data values

fromsklearn.preprocessing import LabelEncoder, MinMaxScaler

lableencoder_Y=LabelEncoder()

data.iloc[:,6]=lableencoder_Y.fit_transform(data.iloc[:,6].values)

data.iloc[:,6]

#handle missing values using simple imputer method

fromsklearn.impute import SimpleImputer

imp=SimpleImputer(missing_values=np.nan,strategy='median')

data['Bp']=imp.fit_transform(data[['Bp']])

data['Ane']=imp.fit_transform(data[['Ane']])

data['Sg']=imp.fit_transform(data[['Sg']])

data['Chl']=imp.fit_transform(data[['Chl']])

data['Sod']=imp.fit_transform(data[['Sod']])

data['Pot']=imp.fit_transform(data[['Pot']])

data['Bun']=imp.fit_transform(data[['Bun']])

data['Scr']=imp.fit_transform(data[['Scr']])

data['Hgb']=imp.fit_transform(data[['Hgb']])

data['Rbcc']=imp.fit_transform(data[['Rbcc']])

data['Wbcc']=imp.fit_transform(data[['Wbcc']])

data['Mcv']=imp.fit_transform(data[['Mcv']])
```

```
data['Pltc']=imp.fit_transform(data[['Pltc']])

data['Pcv']=imp.fit_transform(data[['Pcv']])

#Feature scaling

#min-max scalar method scaled the data set so that all the input features between 0 and 1

data["Bp"]=data["Bp"].apply(lambda v:(v-data["Bp"].min())/(data["Bp"].max()-data["Bp"].min()))

data["Sg"]=data["Sg"].apply(lambda v:(v - data["Sg"].min())/(data["Sg"].max()-data["Sg"].min()))

data["Chl"]=data["Chl"].apply(lambda v:(v - data["Chl"].min())/(data["Chl"].max()-data["Chl"].min()))

data["Sod"]=data["Sod"].apply(lambda v:(v - data["Sod"].min())/(data["Sod"].max()-data["Sod"].min()))

data["Pcv"]=data["Pcv"].apply(lambda v:(v - data["Pcv"].min())/(data["Pcv"].max()-data["Pcv"].min()))

data["Pot"]=data["Pot"].apply(lambda v:(v - data["Pot"].min())/(data["Pot"].max()-data["Pot"].min()))

data["Scr"]=data["Scr"].apply(lambda v:(v - data["Scr"].min())/(data["Scr"].max()-data["Scr"].min()))

data["Hgb"]=data["Hgb"].apply(lambda v:(v - data["Hgb"].min())/(data["Hgb"].max()-data["Hgb"].min()))

data["Rbcc"]=data["Rbcc"].apply(lambda v:(v - data["Rbcc"].min())/(data["Rbcc"].max()-data["Rbcc"].min()))

data["Wbcc"]=data["Wbcc"].apply(lambda v:(v - data["Wbcc"].min())/(data["Wbcc"].max()-data["Wbcc"].min()))
```

```
data["Mcv"]=data["Mcv"].apply(lambda v:(v - data["Mcv"].min())/(data["Mcv"].max()-data["Mcv"].min()))

data["Pltc"]=data["Pltc"].apply(lambda v:(v - data["Pltc"].min())/(data["Pltc"].max()-data["Pltc"].min()))

#with the following function we select highly correalted features

# it will remove the first feature that is correlated with any thing other feature

def correlation(dataset, threshold):

col_corr=set() #set of all the ames of correlated columns

corr_matrix=dataset.corr()

for i in range(len(corr_matrix.columns)):

for j in range(i):

if abs(corr_matrix.iloc[i,j])>threshold: #we are interested in absolute coeff value

colname=corr_matrix.columns[i] # select the name of columens

col_corr.add(colname)

returncol_corr

#show higly correlated features

corr_features=correlation(X_train, 0.7)

len(set(corr_features))

# drophigly correlated features from the dataset

X_train.drop(corr_features, axis=1)

X_test.drop(corr_features, axis=1)
```

Appendix B.2 Sample source code for Building and Evaluating Models

```
#create the function for the models
def models(X_train,Y_train):

  #Decision Tree
  from sklearn.tree import DecisionTreeClassifier
  tree=DecisionTreeClassifier(criterion='entropy', random_state=0)
  tree.fit(X_train,Y_train)

  #Random Forest Classification
  from sklearn.ensemble import RandomForestClassifier
  forest=RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=0)
  forest.fit(X_train,Y_train)

  #MLP model with MLPClasifier
  from sklearn.neural_network import MLPClassifier
  MLP=MLPClassifier(hidden_layer_sizes=(20,20,20,20,20,20), activation='relu',verbose=True,
max_iter=3000,
          solver='adam', random_state=1,alpha=0.005, learning_rate_init=0.005)
  MLP.fit(X_train,Y_train)

   #SVM classifier
  from sklearn import svm
  svmclas=svm.SVC(kernel='linear', gamma='auto',C=40,random_state=1)
  svmclas.fit(X_train,Y_train)

  #Print the models accuracy on the training data
  print('[0] Decision Tree Classifier Trainig Accuracy :',tree.score(X_train,Y_train))
  print('[1] Random Forest Classifier Trainig Accuracy :',forest.score(X_train,Y_train))
```

```
    print('[2] Multi-Layer Classifier Trainig Accuracy :',MLP.score(X_train,Y_train))
    print('[3]Suport Vector Machine Classifier Trainig Accuracy :',svmclas.score(X_train,Y_train))

    return  tree, forest,MLP, svmclas
#Accuracy, Prececison, Recall,Sensitivity, Specificity
fromsklearn.metrics import classification_report, accuracy_score
cm2=confusion_matrix(Y_test,pre)
TP=cm2[0][0]
TN=cm2[1][1]
FN=cm2[1][0]
FP=cm2[0][1]
print('Testing Accuracy =',(TP+TN)/(TP+TN+FN+FP))
print('Testing Precision =',TP/(TP+FP))
print('Testing Recall =',TP/(TP+FN))
print('Testing Sensitivity =',TP/(TP+FN))
print('Testing Specificity =',TN/(FP+TN))
```